More crucially, it should noted that the parameters proposed in 4.38 do not constitute a solution to the assignment of penultimate stress. While they succeed in correctly assigning stress to a seven-syllable word, stress assignment would be sensitive to the length of the string, with the ratio of alpha and beta being roughly equal to the proportion of syllables to the right and left of the assigned stress.\footnote{The question of whether a computational device should formally exclude the possibility of even describing such marked or impossible states (i.e. excessive weak generative capacity) will be addressed in section 6.5.} Generally, analyses where both alpha and beta are positive should be highly marked, if not entirely rejected.

4.3.5 Garawa

Up to this point in the discussion, it might be argued that the computational network proves neither better nor worse as a tool for describing stress systems than competing constituency or grid-based analyses. In large measure, it utilizes parameters that are rather parallel to the parameters described by Halle & Vergnaud or Drescher & Kaye. The description of more complex stress systems such as Garawa or Lenakel provides a better vehicle for comparison. In Garawa, stress is assigned to the initial syllable, the penultimate syllable, and alternating syllables counting back from the penultimate syllable. Additionally, Garawa is stress clash avoidant, preventing a secondary stress from appearing on the second syllable.

For Halle & Vergnaud, the Garawa pattern can be captured by positing the following parameters:

\[(45) \text{Garawa: } [+HT] [+BND] L R/L \text{ (cf. 5c above)}\]
\[
\begin{array}{ccc}
* & \ast & \ast \\
(*) & (*) & (* \ast) (\ast \ast) (\ast \ast)
\end{array}
\]
To account for the pattern in words with an odd number of syllables, Halle & Vergnaud must posit a specific stress deletion rule which they admit "is the only means available to obtain this result." The rule—delete an Line 1 asterisk which is follows a Line 1 asterisk—is arguably, an inelegant and potentially very powerful theoretical device.\textsuperscript{19}

The network, on the other hand, provides a straightforward mechanism for dealing with stress clash. A recognition device which assigns stress to a local maximum directly excludes stress clash by forcing a choice between the competing stresses. The choice of parameters within the system determine which of the two competing stresses will be realized in any given language (Figure 4.46 illustrates odd-parity words; 4.47 illustrates even even-parity words).

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure4.46}
\caption{Figure 4.46}
\end{figure}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure4.47}
\caption{Figure 4.47}
\end{figure}

4.3.6 Lenakel

Lenakel poses an even more difficult set of problems for traditional constituency-sensitive analyses. Halle & Vergnaud describe the distribution of stresses as follows:

In Lenakel, main stress is located on the penultimate syllable in the large majority of words and on the final syllable in a class of specially marked words. As in a great many other languages, the main word stress is preceded by a series of subsidiary stresses. In nouns these fall on every even-numbered syllable preceding the main stress; in verbs they fall on odd-numbered syllables preceding the main stress, except for the syllable immediately preceding main stress (p. 216).

\textsuperscript{19}It is also important to note that the destressing rule posed here is not equivalent to P9 in Dresher & Kaye (cf. #7). Dresher & Kaye's parameter alternatively defoets a weak foot in clash. In the Garawa data, the weak foot receives the stress.
Ignoring for a moment the marked class with word-final stress the principal complication of the Lenakel data is the fact that nouns and verbs appear to require different stress-assignment procedures.\textsuperscript{20} For Halle & Vergnaud, the data requires cyclic rule application, line conflation, two different stress deletion rules and rules that function variably for nouns and verbs. It is also instructive that a block of three rules (the Alternator) is required to account for the default alternating stress pattern.

The dynamic computational network again provides a straightforward account of the distribution of stresses. The case of nouns proves to be no different than Warao (Figure 4.22). In Figures 4.48 and 4.49, we illustrate the treatment of nouns with either odd or even parity.

![Figure 4.48](image1)

![Figure 4.49](image2)

The only modification necessary for the processing of verbs and adjectives is the addition of a morphologically-marked word-initial activation ($I = .50$), creating an appropriate stress clash resolution in words with odd-parity.

---

\textsuperscript{20}Word-final stress is due to the postulation of morphologically marked tense vowels in word-final position. Such a phenomena can be accounted for directly by positing a positive word-final activation in such cases. Rather than merely representing a stipulation, the representation of weight-sensitive or segment-sensitive phenomena will be represented in just this way in quantity-sensitive systems below. The Lenakel data is further complicated by the possibility that more than one of these morphologically marked tense affixes can appear in succession. In such cases, an analysis similar to Goldsmith's (1991) analysis of cyclic stress in Indonesian can account for the data.
4.3.7 Antepenultimate stress

One the strongest potential arguments for Halle & Vergnaud's constituency-sensitive stress assignment procedure is its ability to account for antepenultimate stress. If one utilizes both final extrametricality and right-to-left left-headed constituents, consistent antepenultimate stress can be generated. Halle & Vergnaud consider it to be a strength of their theory that it can generate single stresses on the first three syllables, the last three syllables, but not on intermediate positions. On first glance, it would appear that the computational network could only generate antepenultimate stresses if it adopted the strategy illustrated in Figures 4.41-44 during the discussion of Lakota ($\alpha$, $\beta > 0$). Unfortunately, such an approach would be even more length-sensitive for antepenultimate stress than for penultimate stress and would invite the charge of excessive generative power.

The issue is not so easily resolved, however. First, the set of parameters available to Halle & Vergnaud do permit them to generate single stresses more than three distant from the edge of the word (cf. #6 above). Moreover, examples of primary surface stress which fall more than three syllables from the edge do appear and are ostensibly accounted for using Halle & Vergnaud's rules and parameters. In the discussion of Winnebago below (4.4.5), we must account for surface forms such as wakirípóroporo where the main stress appears on the fourth syllable (Figure 4.67). On the other hand, the network coefficients can be configured in such a way as to reduce the length-sensitivity noted above. If we assume a positive bias and a positive alpha with an arbitrarily small value, we can configure beta and a negative $F$ in such a way as to create a very stable antepenultimate stress (Figure 4.52). Alternatively, we could replace bias with an initial impulse and then permit beta to be slightly greater than 1.00 (Figure 4.53).
4.4 Quantity-sensitive stress systems

In languages that use information concerning syllable weight to assign stress, we need to assign input activation to syllables other than the first and the last. As discussed in 4.1.2, information concerning syllable weight is directly available from the syllable network, permitting either discreet integer weight-values (moras) or a continuously valued weight parameter. Given the additional input activations on heavy syllables, quantity sensitive stress systems also prove to be unique in the fact that they often permit stress on adjacent segments (e.g. Koya, Eskimo) and in the complex interactions which appear between positionally determined stress and weight-sensitive stress (e.g. Komi, Mongolian).²¹

In formal terms, the assignment of additional activation to heavy syllables is conceptually equivalent to Halle & Vergnaud's practice of assigning additional Line 1 asterisks to heavy syllables. The network distinguishes itself from that approach, however, both by permitting a range of input activation values for heavy syllables and in the way different input activations interact with each other. In addition to discussing relatively straightforward cases like Koya and Eskimo (4.4.1), we will examine the network's treatment of more difficult systems like Komi, Eastern Cheremis, Khalkha Mongolian and Aguacatec Mayan (4.4.2). To complete the discussion of stress systems, we will consider the extremely difficult rule interactions that appear in Winnebago (4.4.3) and Creek (4.4.4). As in the case of Garawa and Lenakel among quantity-insensitive systems, it is in the analysis of problems such as Winnebago epenthesis that the true value of the network approach can be appreciated.

²¹Goldsmith (1992) also discusses the interaction of weight and position in Rotuman and Yapese, demonstrating that the network can account for the distribution of final stresses.
4.4.1 Eskimo / Koya

Perhaps the simplest type of quantity-sensitive stress system is that exemplified in languages like Eskimo, where stress is assigned to the final syllable and to every other heavy syllable in the word (Touretzky & Gupta 1992). In a constituency-sensitive approach, such a pattern is achieved by assigning Line 1 asterisks to each heavy syllable and then constructing unbounded right-headed constituents over the entire string. In the computational network, the same pattern is accomplished by simply assigning the positive input activations to $F$ and to each heavy syllable, with no need for either leftward ($\alpha$) or rightward ($\beta$) excitation/inhibition. The recognition device uses a linear threshold ($\Theta > 0$) to assign individual stresses (Figure 4.54). Since Eskimo permits stress clash, a recognition device that identified local maxima would fail to correctly predict all of the stresses in the string.

The case of Koya is quite similar. Rather than assigning positional stress to the end of the word, Koya assigns stress to the first syllable. As reported by Gupta & Touretzky (1991), Koya assigns a lower degree of stress to heavy syllables than it does to the word-initial syllable.\footnote{Halle & Vergnaud (1987) do not report a distinction between levels of Koya stress, but such a factor would not figure in their analysis in any case.} To the extent that Koya (or some similar language) realizes different degrees of stress, the network could model the pattern by assigning different input activation values to the $I$ and $H$ parameters (Figure 4.55).

![Figure 4.54](image1)

![Figure 4.55](image2)

4.4.2 Komi / Cheremis / Mongolian / Mayan

The set of languages exemplified by Komi, Eastern Cheremis, Khalkha Mongolian and Aguacatec Mayan illustrate an interesting interaction between positional activation and weight-sensitive activation (cf. Halle & Vergnaud 1987; Touretzky & Gupta 1992). In Eskimo and Koya, stress appears on both positionally-dependent and additional heavy
syllables. In the present set of languages, both positional and weight-sensitive input activations are present but only a single stress will be realized in the output. In other words, the various input stresses compete for prominence. In the case of Komi, stress appears on the first heavy syllable or on the last syllable if all of the syllables in the word are light. Eastern Cheremis offers the mirror image situation where stress appears on the last heavy syllable or the first if none of the syllables are heavy. Khalkha Mongolian reverses the default direction of Komi by stressing the first heavy syllable or the first syllable if all are light. Mayan represents the mirror image of Mongolian by stressing the final heavy syllable or the last syllable if none are heavy.

Using Komi as an example, Halle & Vergnaud would account for the distribution of stress by assigning Line 1 asterisks to all heavy syllables and then constructing unbounded right-headed constituents. Line 2 asterisks would then be assigned by creating an unbounded left-headed constituent over the Line 1 asterisks. This would assign the stress to the leftmost heavy syllable unless, of course, there were no heavy syllables in which case the only Line 1 asterisk would appear in word-final position. Line conflation would then apply to eliminate the residual Line 1 stresses.

Within a computational network we can account the distribution of stress in the languages in question given the assumption that the network can determine a global maximum as well as local maxima (cf. 4.15). As illustrated for Komi in Figures 4.56-59, given a variety of configurations of heavy and light syllables, the network correctly identifies the stressed syllable as having the highest final activation (Eastern Cheremis would be represented by the mirror image of each of the examples).

![Figure 4.56](image1.png) ![Figure 4.57](image2.png)

---

23The inclusion of a global maximum as a recognition device option, although hinted at, has been avoided thus far due to the significant increase in computational complexity that it imposes on the network. The simplest architecture for computing a global maximum would be to place an additional layer of units on top of a recognition layer that diagnosed local maxima. Alternatively, we could create a competitive net among the units in the recognition layer so that ultimately only one could be active.
The case of Mongolian, where the default stress in the case where no syllables are heavy is on the initial instead of the final syllable, can also be successfully modeled by the network (even without the identification of a global maximum). The key to the Mongolian case is to make the word-initial activation arbitrarily small so it is only realized in cases where no other activations are present in the network (Figures 4.60-61; Mayan is represented by the mirror image).

4.4.3 Winnebago

On first glance, Winnebago offers a very strong warrant for a metrical analysis that is sensitive to the construction of metrical constituents. According to Hale and White Eagle (1980), stress is assigned in Winnebago to every odd-numbered mora in the word except for the first (e.g. 3, 5, 7 etc.). Halle & Vergnaud account for the stress pattern by marking the first mora as extrametrical and then constructing right-headed binary
constituents from left to right. While such a procedure successfully assigns stress to every odd-numbered mora other than the first, it also assigns stress to the final mora in words with an even number of moras. In general, defective binary constituents are assigned a line 1 asterisk. In situations where the directionality of constituent construction and headedness agree this creates no problem.

\[
\begin{array}{cccc}
L/R & Left-headed & & R/L & Right-headed \\
* & * & * & * & * \\
(1) & (2) & (3) & (4) & (5) & (1) & (2) & (3) & (4) & (5)
\end{array}
\]

The binary constituent construction algorithm (the Alternator) overgenerates stress because Winnebago exemplifies a situation where the directionality parameters don't agree.

\[
\begin{array}{cccc}
L/R & Right-headed \\
* & * & * & * \\
<1> & (2) & (3) & (4) & (5) & (6)
\end{array}
\]

Halle & Vergnaud eliminate the inappropriate stress by a rule that deletes an asterisk if it is directly preceded by an asterisk (cf. Garawa stress clash elimination).

A more remarkable justification for constituency comes in the complex interaction between stress assignment and the insertion of epenthetic vowels (Dorsey's Law). Three classes of words can be identified (data from Halle & Vergnaud, p. 32; epenthetic vowels indicated by parentheses).

(62)a. hoshwazhá → hosh(a)wazhá
\(<1> (2 3) <1> x (2 3)\)

ha rakishrujikshnà → ha rakish(u)rujiksh(a)nà
\(<1>(2 3) (4 5) (6) <1>(2 3) x (4 5) y (6)\)

(62)b. ma ashrách → ma ash(á)rach → ma ash(á)rach
\(<1>(2 3) <1>(2 x 3) <1>(2 x) (3)\)

hi rakróbó → hi rak(ó)rohó → hi rak(ó)rohó
\(<1>(2 3)(4) <1>(1 x 3)(4) <1>(2 x)(3 4)\)

hi rakróhoni rá → hi rak(ó)rohóni rá → hi rak(ó)rohóni rá
\(<1>(2 3)(4 5)(6) <1>(2 x 3)(4 5)(6) <1>(2 x)(3 4)(5 6)\)
(62)c. $\text{wakrip\text{r}}\text{ás} \rightarrow \text{wak(i)rip(\text{á}r)}\text{as} \rightarrow \text{wak(i)rip(\text{á}r)}\text{as}$

\[ <1>(2 \ 3) \quad <1> \times (2 \ y \ 3) \quad <1> \times (2 \ y)(3) \]

$\text{wakrip\text{r}óprò} \rightarrow \text{wak(i)rip(ò)rop(ò)ro} \rightarrow \text{wak(i)rip(ò)rop(ò)ro}$

\[ <1>(2 \ 3 \ (4) \quad <1> \times (2 \ y \ 3) \ z \ (4) \quad <1> \times (2 \ y)(3 \ z)(4) \]

In the first group of forms (62a), the epenthetic vowels neither attract stress nor influence the original stress pattern of the word. In the second group, the epenthetic vowels both attract stress and affect the stress pattern to the right of the inserted vowel. In the final group, some of the epenthetic vowels are stressed while others are not. The key observation is the fact that epenthetic vowels that are active appear *within* metrical constituents while the inert vowels appear *between* previously constructed constituents. Since constituents are binary, the insertion of an additional syllable violates the licensing conditions that created the constituent. Halle & Vergnaud then propose the Domino Condition to appropriately restructure the string.

(63) *Domino Condition*: The introduction of an additional position inside a bounded constituent destroys that constituent and all constituents to its right if the Constituent Construction rule applied from left to right, and all constituents to its left if the Constituent Construction rule applied from right to left. Constituent structure is reimposed on the affected substring by a subsequent reapplication of the Constituent Construction rule (p. 33).

Since the application of the Domino Condition is sensitive to whether epenthetic vowels occur within or between previously constructed constituents it is not simply the case that metrical structure is assigned to surface strings. As a result, the Winnebago facts appear to warrant both metrical constituents and ordered derivations.

Before proposing a network alternative, several observations need to be made concerning Halle & Vergnaud's analysis. While it does generate appropriate surface strings, it does so at the cost of a great deal of theoretical machinery and derivational complexity (rule ordering, multiple reapplication of metrical constituent construction rules, stress deletion, extrametricality, etc.). The analysis also results in a great deal of surface opacity. While the insertion of inert epenthetic vowels between constituents could be viewed as "preserving" the underlying stress pattern, it creates patterns that violate surface generalizations. Stress ends up appearing on even-numbered syllables instead of odd-numbered syllables. Words such as *hoshawazhá* begin with three unaccented syllables with a single accent on the fourth syllable. Other words (e.g. *harakishurujikshànà*) have two successive unaccented syllables intervening between accents, thereby apparently violating rhythmicity. On the other hand, the words which contain only active epenthetic vowels (thereby triggering the Domino Condition) appear to satisfy all surface generalizations.

Additionally, the rule interactions between the Domino Condition and epenthesis are not clearly identified. Halle & Vergnaud do not indicate whether all epenthetic vowels are added at one time prior to the application of the Domino Condition or whether each
individual application of epenthesis automatically triggers the Domino Condition if its conditions are met. 24 Rather than being simply an issue for theoretical speculation, the choice potentially has empirical consequences. Given a string a syllables where epenthesis occurs between the second and third syllable and between the fourth and fifth syllable, the variable ordering of epenthesis and the Domino Condition produces different surface configurations.

Extrinsic ordering (stress on x, 4, 5, 7)

Epenthesis  
<1> (2 x 3) (4 y 5) (6 7)

Domino Condition  
<1> (2 x) (3 4) (y 5) (6 7)

Continuous application (stress on x, 4, 6)

Epenthesis  
<1> (2 x 3) (4 5) (6 7)

Domino Condition  
<1> (2 x) (3 4) (5 6) (7)

Epenthesis  
<1> (2 x) (3 4) y (5 6) (7)

Domino Condition  
doesn't apply

Stress deletion  
<1> (2 x) (3 4) y (5 6) (7)

Halle & Vergnaud include a case where the application of the Domino Condition incorporates a syllable which would have appeared between constituents (wakripópro → wakripóropóro), but they do not include the converse, so one could only speculate concerning its treatment.

It should also be noted that the formulation of the Domino Condition results not so much from the natural constraints of the theory as from analytical choices. For instance, one could imagine an analysis which permits foot extrametricality or which posits a rule that destresses the first foot. 25 In such an analysis, left-headed constituents would be constructed left to right beginning with the first syllable. With left-headed binary constituents, we would generate the correct pattern of stresses in words with an odd and even number of syllables without the stress deletion rule that eliminates the stress in the final defective foot in a word with even parity. Of course, we would also destroy the generalization on which the Domino Condition is predicated. Inert epenthetic vowels would appear within constituents and active epenthetic vowels would appear between

24 A similar debate appears in the syllabification literature between approaches the extrinsically order syllabification as a rule within phonological derivations (Kiparsky xxx) and those that permit it to automatically apply every time a phonological rule potentially demands it (Itô xxx).

25 Halle & Vergnaud argue that permitting more than one syllable to be extrametrical would inappropriately give a theory too much power, but they already find it appropriate to permit either syllable extrametricality or segment extrametricality (i.e. word-initial V syllables would be extrametrical but word-initial CV syllables would not be, cf. Archangeli's (1986) analysis of Western Aranda). While foot extrametricality or a rule that simply destresses the first foot may well inappropriately add to the power of the theory, the alternatives required to accommodate the facts also serve to reduce the theory's natural constraints (stress deletion, Domino Condition).
them. In such an alternative world, however, this failure would not be considered an
embarrassment. One would simply reformulate the Domino Condition and justify it with
the observation that all syllables must be incorporated into well-formed metrical structure
(parallel to syllable tactics). Once a licensed foot is constructed it is not sensitive to the
insertion of additional phonological material within it. New material that falls between
well-formed feet would violate metrical tactics, however, triggering the newly
reformulated Domino Condition.

More importantly, Halle & Vergnaud's suggestion that only the construction of
metrical constituents can capture the contrast between active and inert epenthetic vowels
misses a very important observation. With right-headed binary constituents, epenthetic
vowels which lie between constituents follow metrical heads (stress peaks/strong
positions). Those that appear inside of constituents follow non-heads (weak positions).
As a result, a grid analysis could be offered that did not require recourse to metrical
constituents. More importantly, as will be noted below, the distinction between following
a stress peak versus following a trough allows a straightforward network account of the
distribution of stress in Winnebago that eliminates the need to construct (and reconstruct)
constituents.

Before considering the effect of epenthetic vowels on Winnebago stress
assignment, we will briefly consider the basic stress assignment procedure. In 4.3.7 we
described a procedure whereby antepenultimate stress can be modeled by using
appropriate positive values for alpha and beta and a negative final activation. Such a
procedure only generates a single stress, however, rather than the rhythmic pattern
required for Winnebago. The appropriate pattern can be generated, however, with the
following general parameters:26

\[(64) \quad \text{Positional activation: } I > 0; \quad F = 0\]
\[\text{Weight activation: not applicable; } L = 0\]
\[\text{Lexical/Morphological activation: } M > 0 \text{ (assigned to epenthetic syllables)}\]
\[\text{Lateral inhibition: } \beta < -1; \quad \alpha = 0\]
\[\text{Stress assignment: } \theta > I\]

26 In order to account for all Winnebago forms, F must also have a positive value greater than \(|I\alpha| + \theta\).
This will account for bisyllabic forms where the final syllable is stressed. Even though lateral inhibition
from the nominally active initial syllable would make the second/final syllable negative, its own activation
would succeed in making it exceed the threshold. But in longer words with an even-number of syllables,
the cumulative impact of lateral inhibition would outweigh the final syllable's internal activation, thereby
preventing it from being a peak.
The two key parameters that distinguish the system are the values for \( \beta \) and \( \theta \). The use of a value of beta that exceeds -1 in magnitude permits successive wave peaks to increase in value. When combined with a threshold that says a realizable stress peak must exceed the value of the Initial parameter, the parameters create an alternating wave with peaks on odd-numbered syllables. Since the first unit has the value 1 it cannot, by definition exceed the threshold. In effect, the entire first "foot" becomes extrametrical.\(^{27}\)

When epenthetic vowels are inserted into various positions in the string, the network can continue to correctly distribute stress by considering each syllable with an epenthetic vowel to be phonologically active (i.e. possess positive input activation \( M \)). Given an appropriate input value, when a syllable with an epenthetic vowel is inserted after a stress peak, it interrupts the rhythmic structure of the word. Serving as neither a stress peak or trough it, in essence, restarts the wave. While it fails to be a peak both because it is not a maximum and because it falls below the threshold, it generates a new wave that causes the syllable two distant to become a stress peak (Figure 4.67). While the effect is identical to that generated by Hale & White Eagle's analysis, the mechanism is quite different. Rather than assuming that epenthetic vowels "between constituents" are

\(^{27}\)While such a move might appear to increase the power of the network it does so in a very natural fashion. Each of the parameters that are used are independently necessary for even simple stress systems. Allowing \( \beta \) to have a magnitude greater than -1 falls within the convergence zone of the chart at the beginning of the chapter (\( \alpha \) is 0 or very small). It will also be necessary for quantity-sensitive systems such as Creek (cf. 4.4.4). The stipulation that the value of \( \theta \) exceed the value of \( I \) is only marginally more difficult to learn than learning a fixed value. It is also analogous to analyses that set the value of \( \theta \) as essentially equal to \( I \) or \( I \) multiplied by \( \alpha \) or \( \beta \) (cf. 4.12a, 4.12b, 4.16a, 4.16b). Assuming that fears about excessive generative power justify the rejection of a theory, it should be noted that the linkage between \( I \) and \( \theta \) does not produce such power. It would be computationally very difficult, if not impossible, to set \( \theta \) such that an arbitrary number of stress peaks would either be accepted or rejected. As discussed, it can only to be used to either include only the peripheral stress peaks or exclude only the peripheral stress peaks. Compared to the alternatives offered by Halle & Vergnaud (e.g. segment extrametricality, syllable extrametricality, stress deletion, line conflation), the network model is highly constrained.
inert, it is precisely because they are phonologically active that they can interrupt the rhythmic structure while preserving subsequent stresses.

![Diagram of Winnebago stresses](image)

Figure 4.67

In the case where epenthetic vowels are inserted after a stress trough, the resulting stress configuration is achieved in a straightforward fashion. The introduction of an additional input corresponds to a position that would be a stress peak anyway. The surface result is therefore strictly rhythmic. Since the epenthetic vowel adds a position to the string, each subsequent stress is moved, thereby modeling the effects of the Domino Condition.
In the next illustration, we consider a case where several epenthetic vowels are added to a single word (\textit{wakriprópro} $\rightarrow$ \textit{wak(t)rip(ó)rop(ó)ro}). The network not only succeeds in accommodating several additional inputs, but more importantly, correctly leaves the first three syllables unstressed.
In the proposed analysis, surface stress can be predicted directly from surface syllabification as long as the positions of the epenthetic vowels are marked by the input activations. Depending on where the additional inputs are inserted, they serve as either a harmonic or a counter-harmonic influence. No additional machinery or constraints are necessary. The seeming complexity of the Winnebago data can be accounted for without complex rule interactions or reapplication of metrical structure assignment.

4.4.4 Creek "tonal accent"

The American Indian language Creek (Haas 1977) assigns a tonal accent utilizing nearly all of the major conditions that have been described thus far. It assigns a single cumulative stress determined by both position and weight. In words with no heavy syllables, the tonal accent is assigned to the last even-numbered syllable in the word counting from the beginning (penult in odd-parity words, ultima in even-parity words). In words with heavy syllables, stress is assigned to the last even-numbered syllable counting from the last heavy syllable. While such a stress distribution appears to be rather complex it can be modeled with the following parameters:
(70) Positional activation: $I < 0$; $F = 0$
Weight activation: $H > 0$ (approx. $3 \times I$); $L = 0$
Lateral inhibition: $\alpha = 0$; $\beta < -1$
Stress assignment: Global maximum

Several of the possible configurations deserve discussion. In the cases where we have no heavy syllables, we need to account for the different stress assignment in even and odd-parity words. In the cases of Weri (4.3.2) and Warao (4.3.3) we could account for word-final or penultimate stress by adjusting the value of the $F$ parameter. No such procedure could differentially assign stress to ultima or penultimate based on parity, however. If on the other hand, we observe that the stress appears on the last even-numbered syllable counting from the beginning, we can appropriately place the stress by having a negative initial parameter and a negative beta, thereby placing stresses on every even-numbered syllable. In order to place such a stress exclusively on the last even-numbered syllable, we propose a value of beta with a magnitude marginally greater than -1, so that we would have an increasing wave. The recognition device would then select the global maximum (odd parity, Figure 4.71; even parity, 4.72).

![Figure 4.71](image1)
![Figure 4.72](image2)

The effect of introducing a heavy syllable into the string is illustrated in Figure 4.73 and Figure 4.74, where we correctly observe that the stress falls on the last even-numbered syllable following the heavy syllable. The key is to give heavy syllables a positive activation such that wherever they appear in the string, they act as if they are in the second position. In Figure 4.73 we observe the crucial case where heavy syllable appears in the third position but supplants the second syllable as a local maximum.
The final and most challenging case is one where the string contains adjacent heavy syllables. Figure 4.75 illustrates this final configuration.
4.5 Conclusion

The analyses that have been reported in this chapter by no means exhaust the range of stress phenomena that appear in human languages. Even if all languages that have been tested to date (40 languages listed in footnote 3 above) were exhaustively considered in this discussion, we would still only scratch the surface. The goal of the study has been to rather demonstrate the range of theoretical devices which are available within the network architecture and to demonstrate the ability of the network to account for data whose description would otherwise prove difficult (e.g. Winnebago). The study of stress phenomena, particularly as they relate to other phonological processes which might occur within the syllable or autosegmental networks remains as a promising area for future research.
Chapter 5

Phonological Processes / Phonetic Realization Rules

In chapters three and four we have discussed how a dynamic computational network can account for syllabification and metrical structure assignment. While these processes are critical to phonological description, they constitute only a small subset of the domain of phenomena that are of interest to phonologists. Traditional generative phonologies typically propose an abstract underlying form that is progressively transformed into the surface phonetics through a series of derivationally ordered rules (e.g., insertion, deletion, assimilation, metathesis, etc.). While the theoretical innovations of the past twenty years have greatly increased the descriptive power of phonological representations, thereby reducing the number, complexity and the extrinsic ordering of rules, a modicum of rules that modify representations crucially remain as weapons in the arsenal of the phonologist. The current chapter will add to both the descriptive and explanatory power of these residual rules by a) facilitating the manipulation of scalar variables such as sonority that more perspicuously account for the structural relationships that formal symbolic rules attempt to encode; b) describing a dynamic rather than passive relationship between the elements that enter into the structural description of a rule; and c) justifying the application of rules based on their ability to improve the well-formedness (i.e. harmony, periodicity) of the rule's output.¹

In traditional generative formalisms, the structural description of rules typically consist either of simple strings of feature matrices (1) or more complex geometric structures whose ultimate terminals again consist of feature matrices (2-3).

(1) Generative Phonology (SPE)

Standard rule schema:

\[ A \rightarrow B / X \_ Y \] (where A, B, X, Y stand for feature matrices)

¹While the network appears to hold great promise in the modeling of putative phonological rules, the results reported in the present chapter are only suggestive. A great deal more work will be required before any suggestion could be made that the network will permit us to supplant alternative generative theories in the processing of productive phonological rules.
Transformational rule schema:

*Compensatory Lengthening* (Kisseberth & Kenstowicz 1979)

\[
\begin{align*}
SD & V \ ? \ \{C\} \\
SI & 1 \ \{#\} \\
SC & \begin{bmatrix} 1 & \ \{2\} \\ +long \ \{\emptyset\} \end{bmatrix} \\
\end{align*}
\]

(2) Autosegmental Phonology

*Leftward Spread* (Goldsmith 1992)

\[
\begin{array}{ccccccc}
V & C & V & C & V & C & V \\
\end{array}
\]

(3) Metrical Phonology

*Destressing (foot-based)* (Hogg & McCully 1987)

While each of the above theories place both formal and substantive constraints on what each of the ultimate terminals might be, these atoms possess no intrinsic content. Any legitimate member of the set of terminals can appear in any appropriate position, with evaluative judgments made only on metatheoretical criteria (e.g. symbol counting, *naturalness* criteria, etc.; cf. discussion in *SPE* epilogue). Moreover, the symbols are passive with respect to the rules that operate on them. For example, in the statement of assimilation rules within standard generative phonology, there is no intrinsic reason why a given feature that appears in the structural description of the environment should be copied to the feature matrix of the target segment—other than the fact that the rule is
called *assimilation*. Although autosegmental phonology employs a dynamic set of metaphors (e.g., *spreading, harmony*), its formal apparatus still employs externally stipulated rules. In autosegmental spreading, for instance, association lines are added by the rule. There is no formal sense in which the source *propagates* the lines or the target segments *attract* them, despite the metaphors that are often used to describe the process. A dynamic network with spreading activation provides a much more appropriate vehicle for the metaphors that already describe a large number of phonological processes (*attract, propagate, spread, assimilate, weaken, strengthen, lenite*).

5.1 Sonority-sensitive phonological processes

While the arguments concerning the passivity and lack of intrinsic content for the symbols in generative rules are well-rehearsed, an additional problem appears when we discuss scalar variables such as sonority and multiple degrees of stress. If we determine that sonority is a valuable phonological variable, we need the formal machinery to represent and manipulate it (Clements 1987) since the generative formalisms described above are ill-suited to the task of representing sonority relationships (see below).

In order to justify the value of sonority as a phonological variable, we will first consider rules where the structural descriptions and/or structural changes appear to encode sonority-sensitive relationships. Because the network recasts syllable constituency in terms of sonority relationships, we should expect that phonological rules which appear to be sensitive to syllable constituency (i.e., membership in onset vs. coda) can generally be accounted for directly by the network without intermediate mention of constituency. For instance, a rule that spirantizes intervocalic stops or one that aspirates voiceless stops when then are followed by stressed vowels could be reanalyzed in terms of the sonority relationships that appear between the segments. A rule that has the effect of increasing the sonority of a low-sonority segment \( \ell \rightarrow \theta \) when it is in a particular relationship to neighboring high-sonority segments \( \nu \rightarrow \nu \) can be more effectively processed within a computational network than in a formalism the manipulates feature matrices or autosegmental association lines (cf. 5.1). In this regard, the network proves to be capable of generally accounting for fortition and lenition phenomenon (5.1.1), as well as a wide variety of specific allophonic processes such as aspiration/deaspiration (5.1.2), flapping (5.1.2) and length assignment (5.1.3).

5.1.1 Fortition/Lenition

It has frequently been observed that in the historical development of languages, consonants often tend to undergo the process of lenition (or less frequently fortition). Rather than representing random changes in consonantal quality, segments appear to undergo systematic mutations in the direction of higher sonority. Lass & Anderson (1975) describe the sequence of changes as following (cited in Anderson & Ewen 1987; see also Vennemann 1972; Taylor 1974; Foley 1977; Escure 1977):
(4) Sequence of lenition processes\(^\text{2}\)

a. (intervocalic) voiceless stop → voiced stop
b. voiced stop → voiced fricative
c. voiced fricative → approximant consonant
d. approximant → vowel
e. vowel → Ø

It is probably not coincidental that this strength hierarchy that governs the historical process of lenition (i.e. voiceless stops / (voiced stops or voiceless fricatives) / voiced fricatives / nasals / liquids / glides; cf. Foley 1977), proves to be the mirror image of the sonority hierarchy that governs syllabification (cf. 3.1.2).

As Anderson & Ewen (1977) observe, however, models that simply assign distinctive features to each class of segments entirely fail to capture the hierarchy. Each individual change would have to be represented by a different rule since creating an adequate rule schema would prove unwieldy.

(5) Featural representation of major classes (SPE)

\[
\begin{array}{c|c|c|c|c}
\text{+cons} & \text{+cons} & \text{+cons} & \text{+cons} & \text{+cons} \\
\text{−cont} & \text{+cont} & \text{+cont} & \text{+cont} & \text{+cont} \\
\text{+son} & \text{+son} & \text{−syl} & \text{−syl} & \text{−syl} \\
\end{array}
\]

obsruents  nasals  liquids  glides  vowels

More importantly, the lenition processes that are observed diachronically by Taylor, Lass, Anderson, Vennemann and others also appear synchronically, generally proving sensitive to syllable constituency (Foley 1977). In general, onsets constitute strong positions while codas and intervocalic positions prove to be weak.

In Tiberian Hebrew, for instance, the phonemes b, g, d, k, p, t have two reflexes.

In the environment \{C\} we get the ordinary stop consonants while in the environment V we observe the spirants β (or ν), γ, δ, χ, φ (or f) and θ. The crucial observation is that the segments are realized by their more sonorant counterparts when they follow a vowel. Rather than accounting for this pervasive process by referring to feature matrices

\(^2\)Anderson & Ewen (1975) note elsewhere that steps (a) and (b) are often replaced by an alternate path that goes from (a') voiceless stop → voiceless fricative to (b') voiceless fricative → voiced fricative. A dynamic sonority model by noting that the sonority relationship between voiced stops and voiceless fricatives does not need to be established as part of Universal Grammar.
or syllable structure conditions, the network permits a straightforward description in terms of the inhibitory/excitatory dynamic of the high sonority vowel. If beta is positive, rightward excitation will result in a higher derived sonority for the following stop, resulting in its phonetic realization as a spirant.

The key to the network description of lenition is the comparison of a segment's derived sonority values across a variety of environments. Although we have observed throughout that segments take different derived sonority values depending on their environment, the actual value represented within the lateral inhibition network has done very little work thus far. The difference between inherent sonority and derived sonority has only served in a gross way to diagnose syllable structure anomalies (e.g. English onset /fn/ predicted by inherent sonority to be good but its bad; Polish onset /w/ predicted by inherent sonority to be bad but its good; Berber /tt/ rather than /ttu/). A more provocative possibility is that we can predict not only syllabicity from derived sonority values but some measure of phonetic realization as well.

In the discussion of Hebrew above, we suggested that a positive beta (independently needed for Hebrew syllabification) would result in stops that have a consistently higher derived sonority in lenition environments than elsewhere. The observation of higher derived sonority values can be translated into a prediction for spirantization in a variety of ways. First, it would be possible to test all of the environments in which stops appear and then observe whether a segment exceeded an arbitrary threshold. In the illustration in (6), we construct an arbitrary problem with the following parameters: \( \alpha = 0.0, \beta = 0.2, (S)top = 1, (C)ontinuant = 3, (N)asal = 5, (L)iquid = 7, (V)owel = 10; \) threshold for realization as a continuant - 3.00 (same value as underlying continuants).

\[
(6) \quad \begin{array}{l}
t & 1.00 \text{ (inherent value)} \\
#tX & 1.00 \text{ (no leftward inhibition)} \\
StX & 1.20 \\
CtX & 1.60 \\
NtX & 2.00 \\
LtX & 2.40 \\
VtX & 3.00 \text{ (realized as 0)}
\end{array}
\]

While it would be nice to hope for a situation as simple as the arbitrary problem where the derived sonority of the underlying segment would match the inherent sonority of the surface segment, real phenomena are never so well-behaved. What is more likely is that we can observe relative differences in derived sonority values among the variety of environments, with a correlation between the relative value and the allomorphic variant. The choice between variants can either be categorical with appropriate lines being drawn to distinguish between phonetic reflexes or probabilistic based on position along the cline. Ultimately, it might be possible to assign these relative differences in derived sonority to a unit in autosegmental network whose activation value would predict whether the feature were on or off, but such a program lies beyond the reach of the current study.
To summarize, we can treat lenition and fortition as dynamic processes where a sufficiently positive beta would tend to cause coda lenition, while a negative alpha would result in onset fortition. In languages where both alpha and beta are positive, intervocalic consonants would be influenced by both neighboring vowels to have higher derived sonority, thereby accounting for the most common domain for lenition.

5.1.2 Phonetic realization of English voiceless stops

Within the generative tradition, one of the earliest justifications for the syllable as a phonological constituent flows from the observation that several phonological rules can be greatly simplified by making reference to the syllable in their structural descriptions. Following Hoard (1971) and Bailey (1978), Selkirk (1982) observes that the phonetic realization of voiceless stops in English is dependent on whether the segment is assigned to the onset or the coda. She obtains a maximally general condition for aspiration by limiting it to syllable-initial segments.

(7) *Aspiration* (Selkirk 1982, p. 363)

\[
[ + \text{cons} ] \\
[ - \text{cont} ] \rightarrow [ + \text{aspirated} ]_o (\ldots)_o \\
[ - \text{voice} ]
\]

While the introduction of the category, *syllable*, permits a maximally simple statement of the aspiration rule, the challenge is to correctly parse syllables so that all and only the aspirated allophones of the voiceless stops appear in syllable-initial position (cf. discussion of Icelandic syllabification in chapter three where the correct syllable parse becomes a precondition for length assignment). If one adopts the traditional description of English syllabification where the Maximal Onset Principle preferentially builds the largest legal onsets, several apparent counter-examples to the aspiration rule appear (data from Selkirk 1982).

(8) *wacky, attitude, happen, Hegate, accolade, beaker, goiter*, etc.

To account for the data within generative phonology, one could revise the aspiration rule, revise the general syllabification procedure for English, hypothesize ambisyllabicity for intervocalic consonants (Kahn 1977) or propose resyllabification functioning in an ordered derivation (Selkirk 1982).

Dynamic computational networks offer an additional alternative. While it might be possible to train a network so that it appropriately distributes voiceless stops into onset (left side of wave) or coda (right side of wave) positions, it may also be possible to
directly predict aspiration from the derived sonority values of the voiceless stop. In addition to organizing a sequence of segments into syllables and permitting well-formedness judgments, the derived sonority values might license the presence or absence of non-distinctive features. Following Venneman (1972), we could assume that stops are organized into the following hierarchy in terms of sonority:

(9) \[ \text{voiceless aspirated} \quad \text{voiceless unaspirated} \quad \text{voiced} \quad \text{increasing sonority} \]

In order to predict aspiration from derived activation, we would need to discover a relative difference between \( t'/t \), \( p'/t \) and \( k'/k \) such that voiceless stops in aspiration environments consistently have lower derived activation values than in corresponding non-aspiration environments. A close examination of the aspiration data demonstrates that just such a prediction can be made and that such an analysis proves superior to Selkirk's resyllabification rules.

In order to compare Selkirk's generative analysis to the network alternative, we will identify the key data, evaluate Selkirk's proposal and then describe how the network accounts for the same set of data (all but the bracketed data are from Selkirk, pp. 363-365; \( T \) - voiceless stops; \( O \) - obstruents; \( N \) - nasals; \( R \) - sonorants; \( G \) - glides).

(10)a. \#TV - Toronto, pathetic, calamity
    \#TRV - [train, plane, crane, tropic, plastic, craggy]
    b. VTV' - repair, recant
       VTRV' - atrocious, apply, accretion, betwixt, acquaint
    c. VOTV - elliptical, aptitude, actor, after, ictus, hefty, restrictive,
       productive, napkin, Atkins, Riffkin, Lefkowitz (optional)
       VOTRV - actress
    d. VNTV - contemplate, pantomime, winter, center, wimpy, ampersand,
       contemporary, anchor, lanky, linkage, Inca (optional)
       VNTRV - countrified, implication
    e. VITV - filter, altitude, heller, poltergeist, alcohol, Wilkins (optional)
       VITRV - [Coltrain]
    f. VRvTV - particle, turpitude (distinctly unnatural)
       VRvTRV - [partridge]

\(^3\)The use of the network to assign segments to one side of the wave is uniquely problematic in the case of VCV sequences. Since the network identifies only peaks and valleys, directly distinguishing between V.CV and VC.V is impossible. To the extent that constituents are to be mapped from the sonority wave, single intervocalic consonants have been assigned to onsets by stipulation (in accordance with the Maximal Onset Principle). As noted in 2.4.3, VC.V could be distinguished by creating a coda threshold. In other words, we could specify a threshold above which a segment will be assigned to the coda and below which it is assigned to the onset (or marked as ill-formed). Of course, if we use the relative sonority value of a stop in different environments to assign it to coda vs. onset, we could use the same diagnostic to directly predict aspiration or other phonetic features from the derived activation value.
g. V'TV - happy, mightiest, accolade, beaker, goiter, wacky, attitude
   (distinctly unnatural)
V'TRV - *acrimony, *acclamation, [Oklahoma]
h. #sT(R)V - stir, spinach, skate, string
VsT(R)V - Estelle, despair, askance, destroy, display, discreet
VsT(R)V - westerly, aspen, mistress, miscreant, explicate
i. VtIV - Atlantic, antler, atlas

Excluding the cases where the voiceless stop is word-initial (aspirated) or preceded by /s/ (unaspirated), Selkirk notes that the principal environmental distinction between the aspirated and unaspirated allophones is whether the stop is immediately followed by a stressed or an unstressed vowel. If the stop is followed by an unstressed vowel and preceded by a vowel, resyllabification and, as a result, non-aspiration is obligatory. If such a stop is preceded by an /r/, aspiration is "distinctly unnatural." If the stop is preceded by an obstructive, nasal, or /I/, aspiration is optional. Selkirk accounts for the distribution by a rule schema that collapses two resyllabification rules (for purpose of the rule /r/ but not /l/ is [-cons]).

(11) Resyllabification (Selkirk 1982, p. 366)

\[
\begin{array}{c|c|c|c|c|c|c}
1 & 2 & 3 & \emptyset & 4 & 5 & \Rightarrow \\
1 & 2 & 3 & \emptyset & 4 & 5 \\
\end{array}
\]

While the rule generally accounts for the distribution of aspirated stops it does so at significant cost. While the initial aspiration rule elegantly uses the syllable constituent to simplify its structural description, the rule that resyllabifies the string is very complex. The format of the rule is entirely string-based, referring to both preceding and following segments whether or not they are part of the same constituent.\(^4\) In fact, the description of preceding material (different constituent) may indeed be more crucial than the description of the following material (same constituent). It would be possible to replace the stipulation that the following vowel be unstressed with a stipulation that the preceding vowel be stressed. While the two alternatives are generally equivalent due the preponderance of binary feet in English, the choice between the two can be tested in those cases where both the preceding and following vowels are unstressed. For instance, Selkirk notes in a footnote that the /l/ in editor is often aspirated even though it meets the

\(^4\) One of Selkirk's initial arguments for the internal syllable constituents echoes Pike (1967) and Kurilowicz (1948) in the claim that phonotactic constraints (and by extension, phonological rules), should be more likely to operate within than between constituents.
structural description of the obligatory half of the resyllabification rule. She argues that
the unexpected aspiration appears to be due to the fact the stop is two syllables distant
from the stress. While Selkirk explains the anomaly by an appeal to higher order prosodic
structure constraints (cf. McCarthy 1977), a resyllabification rule that is sensitive to the
stress of the preceding syllable could account for the facts in a more straightforward
fashion.⁵

One of the potentially odd features of the rule is the fact that the [+syll, -stress]
designation for the fourth element automatically excludes all tr, pr, kr, pl, and kl clusters
from the resyllabification rule (Selkirk's data (28d); cf. second lines of (10c-g) above).
Even if we accept her judgments concerning the aspiration of voiceless stops in these
environments, the rule is surprising. The rule suggests that "being followed by a liquid" is
actually a more reliable diagnostic for aspiration than whether the segment is either
preceded or followed by a stressed vowel. The judgments concerning aspiration in these
environments also do not appear to be clear-cut. Just as Selkirk notes a cline in the
environments: V'OTV (aptitude), V'NTV (winter), V'ITV (filter), V'rTV (particle) and
VTV (happy), with each successive example less likely to be aspirated (i.e. the
resyllabification rule is option for the first four and obligatory for the last two); words with
stop-liquid clusters also appear to form a cline: VOTRV (actress), VNTRV (countrified), VITRV (Coltrain), V'rTRV (partridge), and VTRV (acrimony). While
both of the parallel lists systematically vary in the likelihood of aspiration, in each pair the
version with an onset-liquid cluster proves to be more likely aspirated than its simple onset
counterpart. Even so, in the case of the VTRV sequences, my intuition differs from
Selkirk, resulting in both acrimony and acclamation being pronounced without aspiration.

As written, the structural description of the rule also applies more generally than
we would have phonetic evidence to support. While the rule is designed to resyllabify
voiceless stops that appear in the appropriate environments it, in fact, obligatorily
resyllabifies all consonants in the environment VCV (→ VC.V) where the final V is
unstressed and optionally, all VCCV (→ V.C.V). While Selkirk's rule makes the strong
claim that English has VC.V vs. V.CV (VCV may be ambiguous), this makes the
Maximal Onset Principle nearly vacuous, except possibly as a prediction that a majority of
disyllabic words have final stress.

Most importantly, the rule schema misses the crucial observation that
resyllabification (deaspiration) is sensitive to the relative sonority of adjacent segments. If
we assume that stressed vowels have significantly higher derived sonority than unstressed
vowels⁶ and we assume that English has a slightly positive alpha (avg. .037) and a
significantly positive beta (avg. .143), the various environments identified in (10a-g)
produce a cline with respect to the derived sonority of the voiceless stop.⁷ The primary

⁵The case of the /t/ in Hecate is a little less clear. While it seems less likely in my speech to aspirate the
/t/ of Hecate than the /t/ of editor, both seem possible.
⁶The potential influence of stress on sonority is well-established (cf. Wiltshire, 1992).
⁷The attempt to describe the aspiration phenomena in the context of the dynamic computational network
uniquely the advantages of the model (particularly the learning algorithm described in chapter six). In
early descriptions of English, we assumed the primary source of lateral inhibition would be a highly
negative alpha with a slightly positive beta. Since it seemed that the crucial environment for aspiration
diagnostic for aspiration proves to be the relative sonority of the lefthand environment. As the sonority of the lefthand environment increases from (10a) to (10g) the likelihood of aspiration decreases. Moreover, as we contrast the forms with an onset cluster with those with simple onsets (e.g., VTV' vs. VTRV'), the forms with a liquid intervening between the voiceless stop and the following vowel generally display more aspiration. Even though the righthand environment proves to be less important in determining the derived activation of the voiceless stop, the difference between liquids and vowels in that environment does marginally influence its derived sonority.

The lack of aspiration in the sp, st, sk, and tl clusters requires a somewhat different explanation. Under the assumption that aspiration is sensitive first to the sonority of a segment's lefthand neighbor and secondarily to that of its righthand neighbor, we would expect that voiceless stops that follow /s/ would have very low derived sonority and hence be excellent candidates for aspiration. The fact that they are not aspirated would appear to directly invalidate the proposal that aspiration should be predicted from the derived aspiration of the voiceless stops.

Rather than offering evidence for a computational analysis, the /s/-clusters appear on first examination to offer the strongest potential support for a constituency-sensitive treatment of aspiration. Given the generalization that voiceless stops are aspirated in syllable-initial position and not aspirated elsewhere, voiceless continuants preceded by /s/ would not be aspirated. Similarly, the unacceptability of /tl/ onset clusters in English would insure that /t/ would always be syllable-final and hence ineligible for aspiration.

While the behavior of /s/-clusters seems to support a constituency-sensitive analysis, they still create problems for Selkirk when they appear in word-internal positions. She observes that the lack of aspiration in words like westerly, aspen and Tuscaloosa could be accounted for by either suggesting that the resyllabification rule doesn't apply or that it resyllabifies both the /s/ and the voiceless stop. Unfortunately, either revision would further complicate the statement of the already complex resyllabification rule. As the resyllabification rule is formulated by Selkirk, words like westerly and aspen meet the structural description of the optional version of the rule. In order to argue that resyllabification can't take place, Selkirk would have to revise the rule to exclude /s/’s.

was to have a stressed vowel to the right, we assumed that the lower-sonority aspirated stops were generated by right-to-left inhibition. The apparent aspiration in VTRV' and VTRV forms proved problematic, however, particularly under the assumption that the activation/inhibition functions were linear. After developing an automatic training device for the network, we discovered that the average final values for alpha and beta were +.037 and +.143, respectively. While this result was surprising, it provides for straightforward predictions for all the (10a-g) forms.

The only potentially difficult transition appears between (10b) and (10c). If we assume that the forms in (10b) are more likely to be aspirated, it would be necessary to have VO produce a greater excitatory effect than unaccented V. While it might seem implausible that the obstruent following a stressed vowel might have higher derived activation than the shewa in words like {repair}, experimental simulation demonstrate that just such a result can successfully be modeled.
(12a) Resyllabification (possible revision 1)

\[
\begin{array}{c}
X \\
\{ [+cons] \\
[-cons] \} \\
\{ +syl \} \\
[-syl] \\
\sim s?? \}
\end{array}
\begin{array}{c}
\begin{array}{c}
[ +syl \] \\
[-syl] \\
_{-stress} \}
\end{array}
\end{array}
Y
\begin{array}{c}
OBL \\
OPT
\end{array}
\]

\[
\begin{array}{lllll}
1 & 2+3 & \emptyset & 4 & 5 \\
\Rightarrow & & & & \\
1 & 2+3 & \emptyset & 4 & 5
\end{array}
\]

A decision to resyllabify the entire cluster would also require a revision that would further complicate the rule. The resyllabification approach would also result in rather surprising syllabifications such as west.er.ly and asp.en.

(12b) Resyllabification (possible revision - Selkirk, p. 369)

\[
\begin{array}{c}
X \\
\{ [+cons] \\
[-cons] \} \\
\{ +syl \} \\
(s)[-syl] \\
_{-stress} \}
\end{array}
\begin{array}{c}
\{ +syl \} \\
_{-syl} \}
\end{array}
Y
\begin{array}{c}
OBL \\
OPT
\end{array}
\]

\[
\begin{array}{lllll}
1 & 2 & 3 & 4 & 5 \\
\Rightarrow & & & & \\
1 & 2+3 & \emptyset & 4 & 5
\end{array}
\]

While the syllable-initial condition for aspiration would appear to correctly block aspiration in sp, st, and sk, an idiosyncrasy in Selkirk's approach makes even this result somewhat doubtful. Elsewhere in her study of English syllabification, Selkirk argues that the clusters in question fill a single syllable-initial slot in the English syllable template (necessary to account for the apparent violation of the sonority hierarchy and the general constraint that the second element in the English syllable template must be [+son]; cf. 3.4.2). In one case, Selkirk's analysis is forced to split the segments and treat them as if they fill two slots and in the other case, they are treated as if they fill a single slot.

Returning to the network account, it is possible that the failure to predict the lack of aspiration in /s/-clusters is due to incorrect assumptions concerning the segments that are input into the network. In each case, we have assumed that the segment in question is a voiceless stop. In the case of st, sp, and sk, however, it is important to note that the environment not only excludes voiceless aspirated stops but voiced stops as well. While little justification could be adduced for arguing that the segments in question should be underlyingly voiced, suggesting that they are initially unmarked for voicing could answer several questions concerning their distribution. If /p,t,k/ were unmarked for voicing (i.e.

---

9The case of /th/ will be addressed in 5.1.3.
/p,t,k/ we would then have to ask what effect that fact would have on the inherent and derived sonority of the segments. In general, underspecified segments exhibit higher inherent sonority than their fully specified counterparts.\textsuperscript{10} Since the network effects are only excitatory ($\alpha > 0; \beta > 0$), any inherent sonority value for /p,t,k/ greater than or equal to the derived sonority threshold of the voiceless unaspirated allophone would insinuate that their derived sonority would also exceed that threshold. Moreover, rather than representing a kluge or stipulation, the present solution actually represents a conservative move, taking an agnostic view concerning the phonemes that underlie /p,t,k/, at least with respect to their inherent sonority.\textsuperscript{11} In other words, with respect to inherent sonority values, we have some segments that are underlyingly equal to the derived sonority of voiceless aspirated segments and others that are greater than or equal to the derived sonority of voiceless unaspirated segments (i.e. /s/-clusters). Within the network, the effect of high sonority left-neighbors (e.g. stressed vowels) is that some of the segments that have inherent sonority within the range of aspirated segments have their derived sonority increased to the range of unaspirated segments (a deaspiration rule in the context of constituency-sensitive grammars). In these terms, instead of aspiration being a fortition process that occurs variably within English, we have a lenition process that functions in a similar fashion to the one described for Hebrew above.\textsuperscript{12}

Before leaving the discussion of aspiration/deaspiration in English, it is imperative that we again address the data reported in (10). In the absence of spectrographic measurements of the voicing onset times for each of the words, we must accept Selkirk's (and in some cases my own) intuitions concerning the absence, presence or degree of aspiration that we find. Judgments range from "obligatory" to "distinctly unnatural" with several gradations in between. When we place these judgments on a cline, many readers may object that the intuitions reported do not match their own. Rather than serving as an

\textsuperscript{10}In a distributed representation of features, each feature can contribute either positively or negatively to the sonority vector (see chapter two, \#8). Features representing point of articulation tend to contribute negatively to the sonority vector so that if the features were not active the sonority vector would have a higher value. In similar fashion, if the feature relating to [voice] were configured as [-voice] rather than [+voice], its failure to be activated would also increase the effective sonority of the segment. Since traditional feature accounts prejudice the issue by pre-selecting the valence of a feature, an appropriately configured autosegmental network should include separate units for both the positive and negative valence of potentially equipollent feature (at least during training).

\textsuperscript{11}Even within constituency-sensitive accounts, the observation that /p,t,k/ might underspecified with respect to the feature [voice] has beneficial results. Their lack of aspiration would be due to the fact that they don't meet the structural description of the aspiration rule (of course, this assumes that the aspiration rule, arguably a late one, would precede the rule that fills out the feature matrix of the ultimately voiceless stops). Concerns about violations of the sonority sequencing principles could also be answered with a caveat that places underspecified segments at a different place in the hierarchy than their fully specified counterparts.

\textsuperscript{12}While the logic of the current approach may seem to be the opposite of traditional accounts of aspiration, it, in fact, mirrors Selkirk's procedure. We are faced with a categorical rule (or constraint) that aspirates syllable-initial segments. The general procedure for forming syllables places all voiceless segments except /l/ and /s/-clusters in this environment, however. As a result, we need a repair strategy to take an appropriate subset of such segments out of the obligatory aspiration environment, thereby having the effect of failing to aspirate or alternatively, deaspirating that class of segments.
embarrassment for the network approach, however, these variable judgments confirm its value. Rather than being faced with categorical rules that either obligatorily or optionally aspirate or deaspirate a segment, the network allows a variety of gradient phenomena and variable judgments that can vary across words, across classes of segments, across speakers and across time.

5.1.3 Icelandic vowel length revisited

In chapter three (3.3) we demonstrated that the network could appropriately account for the idiosyncratic distribution of intervocalic clusters in Icelandic. In that demonstration we proceeded on the assumption that the appropriate syllabification of intervocalic clusters is essential to the correct description of Icelandic phonology. Based on Vennemann's (1972) observation that correlates open syllables and long vowels, appropriate syllabification provides a constituency explanation for length. Itô makes this correlation more natural by describing the correlation as a templatic condition that marks the coda position in the syllable template as obligatory. If the position is not filled by a consonantal segment (due to the Universal Core Syllable Condition (V.CV) or the Icelandic Tautosyllabic Condition (V.trV)), the vowel is lengthened automatically (cf. Maddieson 1985).13

Apart from the mandate to assign segments to onsets or codas in order to account for length, there is little independent evidence for the proposed parsing of intervocalic clusters (cf. Orešnik & Pétursson 1978). If a procedure would be available which could correctly account for the length of Icelandic vowels without recourse to the constituency-sensitive conditions noted by Vennemann, Itô, and others, such a procedure would be arguably superior.

Fortunately, the same computational network that allows us to correctly parse intervocalic clusters in Icelandic provides additional information that may allow us to predict vowel length in initial stressed syllables without recourse to the constituency of the following consonant (onset vs. coda). The recognition layer diagnoses syllable structure by noting peaks and troughs in the derived sonority wave. The actual derived sonority values are not important, however, except in cases where segments must meet a threshold in order to be licensed as a syllable nucleus (e.g. English doesn't license the sonority peak in /pnp/ as a nucleus because it is insufficiently sonorous). Segments do take different derived sonority values, however, depending on their environment. For instance, if \( \alpha < 0 \), a segment will have a higher derived sonority if followed by a low-sonority segment than if followed by a high-sonority segment. In a recurrent network (or in a feedforward network with more than one alpha and/or beta connection), a segment would also be sensitive to the sonority of its more distant neighbors. In a less dramatic fashion (\( \alpha^2 \) in a recurrent network), a segment will have a higher derived sonority if the segment two segments to its right has high sonority than if it has low sonority.

---

13In moraic phonology, this principle can be described as the Principle of Moraic Preservation (cf. Clements 1986).
(13) If $\alpha < 0$ in $VX_1X_2$ the marginal derived sonority of $V$ is inversely proportional to $X_1$ ($\alpha$) and directly proportional to $X_2$ ($\alpha^2$).

If the derived sonority of a vowel is roughly correlated with its length, the distribution of facts described in (13) precisely describes the distribution of long vowels in Icelandic. In chapter three, it was noted that the "correct" syllabification of Icelandic could be achieved with both positive and negative values of alpha as long as beta is appropriately positive (Figure 5.13, same as Figure 3.27).

![Figure 5.13](image)

In those solutions in Figure 5.13 where $\alpha < 0$, the network generally assigns higher sonority to vowels in "open" syllables than it does to the same vowel in closed syllables. The correlation between derived sonority and length is not a linear one, however. It is certainly not the case that long vowels are 50% or even 10% more sonorous than short vowels. We rather observe a threshold above which a vowel can be interpreted as long and below which a vowel is interpreted as short. Alternatively, we could view the excess derived sonority as licensing a vowel length feature.\(^{14}\)

In order to test the possibility of predicting length directly from the network without intermediate reference to constituency, we conducted two tests on an artificial corpus of twenty-eight forms that reflect all of the crucial consonantal contrasts. In addition to the twelve V.CCV forms representing the intersection of \{p,t,k,s\} and \{r,v,j\}, sixteen VC.CV forms with clusters having a slighter smaller sonority difference were

\(^{14}\)One of the potential advantages of the network is to use feedback connections between the syllable network and the autosegmental network to permit derived sonority values to differentially license the ultimate presence or absence of features in the surface phonetics, just as feedback connections between the metrical network and the syllable network can influence the licensing of features in stressed or unstressed syllables.
included in a test corpus, thereby insuring that the network wouldn't overgenerate onsets. In the first test, we simply used the final values for alpha, beta and all of the sonority coefficients that had been trained in the syllabification experiment discussed in 3.3.4. Since that experiment was replicated 175 times, we simply selected the values from the final replication rather than searching for optimally descriptive values (α = −.172, β = .243). Without modifying weights, each of the 28 forms in the test corpus were presented to the network. In (14a,b) we report the syllabification generated by the network for each form as well as the derived sonority values for each segment in the string.

(14a) V: .CCV forms:

e:.pra 9.95 3.01 3.77 11.26
e:.pja 9.98 2.86 4.60 11.47
e:.pva 9.94 3.09 3.29 11.15
e:.tra 10.10 2.15 3.56 11.26
e:.tja 10.12 2.01 4.40 11.47
e:.tva 10.08 2.24 3.08 11.15
e:.kra 9.98 2.86 3.73 11.26
e:.kja 10.00 2.72 4.57 11.47
e:.kva 9.96 2.94 3.25 11.15
e:.sra 10.08 2.24 3.58 11.26
e:.sja 10.11 2.09 4.42 11.47
e:.sva 10.07 2.32 3.10 11.15

(14b) VC.CV forms

ep.la 9.91 3.22 2.50 10.96
et.la 10.06 2.37 2.30 10.96
ek.la 9.94 3.08 2.47 10.96
es.la 10.05 2.46 2.32 10.96
eb.ra 9.52 5.49 4.37 11.26
eb.ja 9.55 5.35 5.21 11.47
eb.va 9.51 5.58 3.89 11.15
ed.ra 9.50 5.66 4.41 11.26
ed.ja 9.52 5.51 5.25 11.47
ed.va 9.48 5.74 3.93 11.15
eg.ra 9.20 7.35 4.83 11.26
eg.ja 9.23 7.21 5.66 11.47
eg.va 9.19 7.44 4.34 11.15
eb.la 9.49 5.71 3.11 10.96
ed.la 9.46 5.88 3.15 10.96
eg.la 9.17 7.57 3.56 10.96
Two observations can be made concerning the output of the network. First, if we assume that Itô's syllabification principles are correct, each of the above forms are "correctly" syllabified (the onset of the second syllable is indicated by the sonority minimum). More importantly, without any attempt to train the network to generate a length distinction, a general distinction can be observed. If all of the forms are listed in descending order based on the derived sonority value of the first vowel, the V.CCV forms would be at the top of the list and the VC.CV forms would be at the bottom. The only exceptions would be et .la and es .la, where the derived sonority of the /e/ exceeds that of several of the V.CCV forms and perhaps ek .la, where the /e/ matches the derived sonority of the /e/ in e .pva.

In the previous test, the derived sonority distinctions between vowels in open syllables and vowels in closed syllables were simply observed as the accidental byproduct of an earlier syllabification experiment. In a second test, we attempted to study the correlation more explicitly. This was done by performing a learning simulation on the 28 constructed forms. Beginning with random values for the sonority coefficients we used the learning algorithm described in chapter six to train descriptively adequate sonority values for each segment in the phonological inventory of Icelandic. While the sonority coefficients were modifiable, we clamped the values of alpha and beta in order to test the hypothesis that the length distinction is sensitive to a negative alpha (α = −.25, β = 0.0). The following results indicate the network's success in predicting length:

(15a) V. CCV forms:

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>e:</td>
<td>.pra</td>
<td>10.02</td>
<td>0.53</td>
</tr>
<tr>
<td>e:</td>
<td>.pja</td>
<td>10.05</td>
<td>0.44</td>
</tr>
<tr>
<td>e:</td>
<td>.pva</td>
<td>9.994</td>
<td>0.65</td>
</tr>
<tr>
<td>e:</td>
<td>.tra</td>
<td>10.18</td>
<td>-0.11</td>
</tr>
<tr>
<td>e:</td>
<td>.tja</td>
<td>10.21</td>
<td>-0.20</td>
</tr>
<tr>
<td>e:</td>
<td>.tva</td>
<td>10.15</td>
<td>0.01</td>
</tr>
<tr>
<td>e:</td>
<td>.kra</td>
<td>10.05</td>
<td>0.41</td>
</tr>
<tr>
<td>e:</td>
<td>.kja</td>
<td>10.08</td>
<td>0.32</td>
</tr>
<tr>
<td>e:</td>
<td>.kva</td>
<td>10.02</td>
<td>0.53</td>
</tr>
<tr>
<td>e:</td>
<td>.sra</td>
<td>10.13</td>
<td>0.10</td>
</tr>
<tr>
<td>e:</td>
<td>.sj a</td>
<td>10.15</td>
<td>0.01</td>
</tr>
<tr>
<td>e:</td>
<td>.sva</td>
<td>10.10</td>
<td>0.22</td>
</tr>
</tbody>
</table>

(15b) VC.CV forms:

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ep .la</td>
<td>9.87</td>
<td>1.15</td>
<td>0.23</td>
</tr>
<tr>
<td>et .la</td>
<td>10.03</td>
<td>0.51</td>
<td>0.23</td>
</tr>
<tr>
<td>ek .la</td>
<td>9.90</td>
<td>1.03</td>
<td>0.23</td>
</tr>
<tr>
<td>es .la</td>
<td>9.98</td>
<td>0.72</td>
<td>0.23</td>
</tr>
<tr>
<td>eb .ra</td>
<td>9.33</td>
<td>3.29</td>
<td>2.70</td>
</tr>
<tr>
<td>eb .ja</td>
<td>9.36</td>
<td>3.19</td>
<td>3.07</td>
</tr>
<tr>
<td>eb .va</td>
<td>9.30</td>
<td>3.41</td>
<td>2.22</td>
</tr>
</tbody>
</table>
In this second test, the only exceptional form is et-la which has a higher value than some of the forms in (15a). While this may be troublesome, the cluster tl proves to be unusual in several languages (cf. 3.2.2, (10i) above). For instance, in Harris’ analysis of Spanish syllable structure constraints, tl must be excluded even though it satisfies the Minimum Sonority Distance Principle (presumably because /tl/ and /ll/ share the same point of articulation, cf. Harris 1983). Larson (1990) accounts for this constraint by suggesting that this /tl/ has a higher inherent sonority because it is underspecified for point of articulation, sharing this specification with the following segment (cf. footnote 10). In addition to accounting for the syllable structure constraints in Spanish (and English as well), the same observation would eliminate the exceptional character of et-la in this test.

While it might be claimed that the network can predict vowel length without recourse to syllable constituency, the above analysis actually argues that the two can be correlated. In all of the forms in (15a) and (15b), the network produced the syllabifications predicted by traditional analyses in addition to predicting vowel length. Perhaps a more compelling demonstration would be to train the network to recognize vowel length without concern for syllabification and then see what syllabifications would result. Alternatively, we could attempt to correlate predictions of vowel length with syllable structures that uniformly conform to the Maximal Onset Principle. Without independent evidence concerning Icelandic syllable structure, however, it is more critical to demonstrate that the network can generate traditional syllabifications.

There are, in fact, additional environments where length and traditional syllabifications do not correlate that demonstrate the superiority of the network analysis. Orešnik & Pétursson (1978) note four classes of exceptions to Icelandic length assignment:

\[
\begin{array}{cccc}
(16) & \text{CV:CC} & \text{CV:C.CV} & \text{CV:Cr} & \text{CV:Cr} \\
\text{ski:ps} & \text{li:t.ka} & \text{pu:kr} & \text{klífr} \\
\text{lei:ks} & \text{no:t.kun} & \text{snu:pr} & \text{grenj} \\
\text{bå:ts} & & & \\
\end{array}
\]

In each of the first two environments, we would expect short rather than long vowels. While final extrametricality can account for the appearance of long vowels in closed monosyllables, the forms in the first column have two final consonants. While the forms are bi-morphemic, an analysis that assigns vowel length before adding the /s/ would
contradict the analysis of inflectional morphology that forms the rationale for either Itô's or Kiparsky's discussion of Icelandic. Even if the forms in the first column were rationalized as morphologically complex, the forms in the second column are not amenable to such an analysis. In both columns, the relevant observation appears to be the fact that the consonant that follows the vowel is from the class \{p,t,k,s\} which typically follow a long vowel. Since the principal determinant of length in the network is the sonority of the following segment, the seeming exceptions are consistent with the network's predictions.

In the final two columns, the gross structure of the forms is identical (CVCr) but the length is variable. Kiparsky (1984) correctly notes that though the forms are analyzed as monosyllables the clusters that follow long vowels are precisely those that would be well-formed onsets in bisyllabic forms. He accounts for the distribution by noting that the forms are derived from verbs ending in /a/. Length assignment must precede derivation after which resyllabification takes place. While possible, such a solution relies on post-lexical derivation, a theoretical move that violates much of the spirit of lexical phonology. The network accounts for the contrast in a straightforward fashion since length is dependent on the sonority profile of the cluster following the vowel rather than syllable structure.

Whether or not the exceptional forms in (16) are admitted into the discussion, the network provides a more direct explanation both for the distribution of long vowels and the syllabification of intervocalic clusters. Rather than positing extra machinery and large-scale differences in directionality to account for Icelandic's unique syllabification phenomena, the network simply assigns marginally different parameters for lateral inhibition.

### 5.2 Additional phonological processes

While several phonological processes can potentially be modeled by reference to derived sonority, a large number of productive phonological processes are not amenable to such an analysis because either the context or the result of the rule proves not to be strictly sonority-sensitive. While lateral inhibition between a unit and its left and right neighbors is a powerful device for describing syllable-sensitive and foot-sensitive phonological processed, it proves to be too blunt a tool to model all of the context-sensitive processes that affect a string. Even permitting an arbitrary number of additional connections between more distant segments in a string, a very costly move in terms of both generative power and computational complexity, would fail to account for processes that are not sensitive to either inherent or derived sonority.

Up to this point in this chapter, we have relied on information that is available within the syllable network (i.e. inherent vs. derived sonority) and top-down information derived from feedback connections between the metrical network and the syllable network (e.g. higher activation for stressed vowels). We have given very little attention to how information is processed within the autosegmental network, however. The autosegmental network has been limited to the role of providing an inherent sonority value to the syllable network, a potentially trivial task that can be performed by a localist network that simply assigns a sonority coefficient to each segment in a phonological inventory. When we
consider phonological processes such as assimilation and harmony, however, we discover a strong justification for a more distributed or featural representation (cf. Chomsky & Halle 1968, pp. 335-340, for a defense of distinctive features in phonological representations).

5.2.1 Assimilation / Dissimilation

One of the most productive rules in the arsenal of generative phonological rules is the process of assimilation, whereby a feature associated with a neighboring segment in the environment is assigned to the target segment (either changing its previous value or filling in a previously underspecified value). While certain assimilations appear to have a sonority-sensitive motivation (e.g., nasal assimilation in coda/onset clusters, complete assimilation/gemination), others appear not to have such a motivation. For instance, several languages exemplify a process whereby /t/ and /d/ alternate with /ç/ and /ø/, where the palatal affricates precede high vowels (e.g. Papago, Kenstowicz & Kisseberth 1979):

\[(17) \begin{bmatrix} \text{stop} \\
\text{dental} \end{bmatrix} \rightarrow \begin{bmatrix} \text{affricate} \\
\text{palatal} \end{bmatrix}_{-} \begin{bmatrix} \text{vowel} \\
\text{high} \end{bmatrix}\]

While any process that makes adjacent segments more alike might arguably increase the harmony of a representation, the rule in (17) can not be explained in terms of differences in derived sonority within the network. Initially it might be tempting to describe the change from dental stop to palatal affricate as an lenition that would be prompted by the high sonority vowel to its right (α > 0). Such a suggestion would present several difficulties, however. We would first have to ask why the lenition would only apply to dental stops rather than across the board since the vowel would have the same excitatory affect on any left-neighbor. Even if this objection could be finessed, the lenition account fails to explain why high vowels trigger he change while other vowels do not. When we observe that it is the high vowels, which by all accounts possess lower inherent sonority, rather than the low or mid vowels, that trigger the palatalization the sonority-sensitive lenition account becomes nearly untenable. One of the principal motivations for lenition as a phonological process and even more crucially, the reason it can be

---

15If we combine the underspecification analysis described in footnote 10 with a stipulation that doubly linked segments in these coda/onset environments assign their featural specification to the second (onset) segment, we provide a straightforward account of coda licensing and Prince languages. The segment in the coda slot is assigned high sonority due to its relative underspecification and, as a result, can be licensed in the coda (cf. Clements & Keyser 1983 where the initial segment fills a V-slot; see also autosegmental licensing account in Goldsmith 1990).

16With a great deal of creativity, even this objection could be finessed. If we assume that alpha is negative and that the underlying segments for /T/ and /D/ are underspecified, we could correlate the highest sonority vowels (low vowels, mid vowels) with the lower sonority stop reflexes for the consonants, and the lower sonority high vowels with the higher sonority palatal affricates. In general, if the sonority of two neighboring segments are inversely correlated, we have evidence for a negative alpha and/or beta.
modeled within the network architecture is the fact that it is an across the board phenomenon. More critically, any attempt to describe the assimilation between dental stops and high vowels as a sonority-sensitive process misses the more likely fact that the high vowels and the palatal affricates appear to share the feature [+high].

While assimilation phenomena can not directly be modeled within the syllable network, it is possible to model them within an appropriately configured autosegmental network (cf. figure 2.11). If the autosegmental network is designed to only compute sonority values each column of units can be isolated from its neighbors. If we permit units (features) in one column to be connected to units in adjacent columns, however, the presence of an active feature in one column could influence its presence in the adjacent column. If the connection weight is positive we would have assimilation while a negative connection would produce dissimilation.

The architecture described in 2.3 proves, in fact, to be a natural implementation of much of the machinery of autosegmental phonology. Feature spreading or the adding of association lines parallels the spreading of activation along positively-weighted connections. Various feature geometries (e.g. Rolodex, class-node, etc.; cf. Goldsmith 1990) could be represented by different patterns of connectivity between columns of features. Underspecification theory could be represented by the fact that units can take positive, negative, or zero inherent activation values, with the possibility of even finer discriminations available with continuous activation values rather than simply +/-0. Finally, such a model could be made truly autosegmental by representing featural units as independently connected to the segmental network rather than appearing in a matrix of identical columns. Just as the interface between the syllable network and the metrical network assigns metrical units to segmental strings of different lengths, autosegmental units could be assigned to segmental strings of variable length. While the parallels between autosegmental phonology and an appropriately configured autosegmental network suggest a productive domain for research, it is a project that remains for future consideration.

5.2.2 Harmony

While the autosegmental module within the dynamic computational network offers the ability to naturally describe assimilation and dissimilation processes (local connections), it proves to be particularly useful in the description of harmony processes (recurrent connections). Just as a single stress can be propagated an arbitrarily long distance within the metrical network as a result of lateral inhibition (perfect grid), the presence of a given feature could be propagated an arbitrarily long distance in the autosegmental network. Whereas a negative alpha and/or beta in the metrical network causes alternating patterns of stress, a positive alpha and/or beta could place a feature on each eligible segment. Patterns of segments that either block harmony or prove to be opaque to harmony processes could potentially be accounted for by controlling not only the magnitude of the coefficient of lateral inhibition/excitation but also controlling the inherent activation of the affected feature for each segment in the string.
5.2.3 Insertion / Deletion / Reduplication

Before leaving the discussion of phonological rules, we need to discuss phonological rules for which we do not yet have an adequate account within the current network architecture. While the syllable network does not impose a great deal of structure on the phonological representation (e.g. no constituency, possibly no stipulated features, etc.), it does represent the string as a linearly ordered sequence of segment-sized units that are only connected to immediate neighbors. Such a model inherently has difficulty with any process that changes the length of the string by either adding or deleting skeletal positions. Once a string is mapped unto the network, even if buffering makes the necessary amount of network rather small (e.g. three positions), it would be difficult to insert a segment between two units. While it would be easier to delete a unit (perhaps by simply making it unrealizable), the deletion would not result in the surrounding units becoming structurally adjacent.

Several possibilities are available for future evaluation. First, it may be that the network's role is simply to parse and/or evaluate the well-formedness of representations that would be generated by a separate production device. Since the dynamic computational network's processing of wave phenomena is not inconsistent with the postulation of a symbolic device that processes particles and constituents, a hybrid model that processes handles length changing processes outside of the network should not be impossible.

Given the widespread desire to present unified theories which handle all processes using basically similar architectures, modifications of the network model should also be considered. Perhaps the fundamental problem is treating time in terms of spatially contiguous units that are both discrete and abstract. The graphs that illustrate the energy waves show the individual units connected by line segments. While these line segments help us visualize the relative magnitude of the various units in the network, they currently possess no reality in and of themselves. No phenomenon is presumed to exist on the line between the two units. The wave metaphor is even more distant from the state of affairs that exists within the network's discreet units. We tend to conceive of harmonic wave phenomena in terms of continuous sinusoidal waves. Given a future architecture that could process time vs. sonority as a truly continuous function, we could imagine stretching the wave both vertically and horizontally, thereby directly modeling a variety on insertion and deletion phenomena (particularly processes such as compensatory lengthening) simply by deforming the energy wave.
Chapter 6
Learning Procedures in Dynamic Computational Networks

Up to this point in our discussion, we have demonstrated that dynamic computational networks can correctly model a wide range of phonological phenomena. Given an appropriate set of coefficients that comprise the knowledge of the system, the network can correctly syllabify, assign stress and model a variety of phonological processes. While the value of the network model is potentially independent of how it acquires its knowledge, one of the greatest potential benefits of connectionist architectures is the existence of powerful automatic learning algorithms that permit a network to induce appropriate connection weights from a limited corpus of data. In the current chapter we will describe the development of training algorithms that can be applied to the task of learning the necessary set of coefficients within the dynamic computational network.

In discussing automatic learning within the dynamic computational network, we intend not only to contribute to the linguistic discussion of language acquisition but to the larger debate in the cognitive science community concerning the value of connectionist networks for the representation of linguistic structure, addressing what Geoffrey Hinton (1991) describes as —

"...the current tension within the artificial intelligence community between advocates of powerful symbolic representations that lack efficient learning procedures and advocates of relatively simple learning procedures that lack the ability to represent complex structures effectively. The contributors (to the 1991 MIT volume, Connectionist Symbol Processing) aim to extend the representational power of connectionist networks without abandoning the automatic learning that makes these networks interesting."

However, even if the proposed learning algorithms are not considered in any way to be psychologically real or an implementation of part of a language acquisition device (LAD), they serve an important heuristic function. In the previous chapters it was suggested that the goal was to select descriptively adequate coefficients. Even when pencil and paper testing is possible, the network provides the opportunity to do in minutes what would otherwise take weeks. In some sense, theoretical linguists historically have often functioned in the same inductive fashion as the learning device. Data-driven linguistics has frequently taken a finite corpus of data, tested its rules, representations, etc. against that data, incrementally revised its theories to deal with a larger percentage of the data, and so on. One of the major differences is that we have attempted to provide a computationally explicit theory that facilitates both the network modeling and, more important, automatic theory testing. Beyond proposing a "new" theory of phonology, the model provides a

---

1Earlier versions of the chapter were presented at Chicago Linguistic Society Parasession on the Cycle in Linguistic Theory in April 1992 and at the Royaumont conference on Compositionality in Neural Networks, July 1992.
procedure whereby a variety of computationally explicit theories can be efficiently tested against realistic corpora of data.

6.1 Methodological Preliminaries

Before describing the development of a learning algorithm, we will briefly describe the data bases and coding conventions used to describe the data. While automatic learning can potentially be applied to each of the types of phenomena discussed in chapters three through five, the following discussion will concentrate on the learning of syllabification systems, both because it has received the greatest amount of attention and because it poses the most difficult learning problem. Several factors make the learning of syllable structure more difficult than the learning of metrical structure (assuming similar descriptive abilities). The syllable network must deal with a larger number of input variables (sonority for each feature of segment as opposed to a small number of parameters -- initial, final, lexical, heavy, bias) and a significantly larger number of output patterns (148 unique syllable patterns in the first 2000 English words as opposed to ~20 unique stress patterns). As a result, most of the ensuing discussion will focus on the development of learning algorithms that discover descriptively adequate values for alpha, beta and the sonority vector such that the network will correctly syllabify any target string.

6.1.1 Data bases

To test the ability of the network to learn to correctly parse syllables in a variety of languages, we have employed three different kinds of data bases.

- High-frequency (relative frequency word lists)
- Problem cases (journal articles, dissertations, etc.)
- Artificial (formal systems, impossible languages, etc.)

Each of the above types offer important insight concerning the operation of the network and its ability to appropriately model linguistic phenomena. Perhaps the most useful but most difficult to create is the natural data base. If one wants to model a procedure by which an intelligent system (e.g. human infant) could learn the network weights, the network presumably would have to be exposed to a similar corpus of data, randomly presented with the same relative token frequency, and coded with the same information that the learner could/would have access to. If one is to use a supervised training regimen where the input is incrementally modified to conform with the target output, the learner would also have to have access to suitably coded output for the training corpus.

While it is nearly impossible to obtain a frequency-weighted corpus representing the speech to which children are exposed, various approximations are available.² Several

²The simulations intentionally use word lists that represent the speech that children are exposed to rather than forms that they use. The choice reflects the judgment that children's ability to recognize syllable peaks and stressed syllables precedes general language production. The training procedure described
researchers have compiled frequency-weighted word lists based on representative text samples. The current study uses four such data bases (the numbers in parentheses indicate the number of words included in the test corpus/total number of words in the published list):

Frequency Dictionary of French Words, Alphonse Juillard, Dorothy Brodin, Catherine Davidovitch, eds. 1970 (620/5082)
Frequency Dictionary of Spanish Words, Alphonse Juillard, E. Chang Rodriguez, eds. 1964 (578/5024)

In addition to objecting that frequency-weighted databases may fail to accurately reflect the speech that children are exposed to, one might argue that they represent too "simple" a corpus. If high-frequency forms expressed fewer irregularities and difficulties than lower-frequency forms, the network might be able to "learn" the high-frequency corpus without being able to process the truly problematic forms. As a result, we also construct databases from the data discussed in the linguistic literature (journal articles, dissertations, etc.). For instance, Dell & Elmedlaouli's analysis of Berber syllabification (1985), Harris' analysis of Spanish (1983), Itô's discussion of Icelandic (1986), and a large French data base supplied by Bernard Laks (p.c.) all serve as important corpora that allow us to evaluate the merits of the network. As noted above, using an automatic learning device allows for a much more efficient testing of these problem cases than traditional paper and pencil methods. While one could not argue that success in learning these restricted databases necessarily represents an ability to learn these languages, they do demonstrate that the network can successfully process difficult syllabification phenomena.

As a final alternative, we can devise artificial databases in order to study the formal properties of the model. For instance, we can create impossible formal languages in order to determine whether the dynamic computational network can indeed learn anything (cf. Fodor & Pylyshyn 1988; see also Prince 1992 for overgeneration charge against Goldsmith 1991, Goldsmith & Larson 1991 models). It is also possible to create artificial databases that permit us to exhaustively analyze a particular phenomenon (cf. Larson & Goldsmith, CLS 1992).

6.1.2 Coding - Input

In order to accommodate a phonological representation to the syllable network, the data must be appropriately coded for both input and output. For input, we adopt a broad phonetic transcription, one that is more abstract than the surface phonetics, yet less abstract than those typically postulated as underlying representations in generative

below also operates passively. The network only need know that its proposed syllabification does not match the target syllabification.
phonological models. While the "failure" to begin with these underlying representations might be viewed critically, several considerations suggest that the forms supplied to the syllable network are appropriate for the task under scrutiny. Even if we agree that underlying representations are justifiably different from surface forms, it is very difficult to determine what the underlying representation for the current model should be. First, it should be noted that the underlying representations postulated by a given analysis are often dictated by theory-internal considerations. In derivational models, underlying forms are often dictated not only by the distribution of surface facts, but also by meta-theoretical concerns such as a quest for generality in rule formalisms, avoidance of rule ordering paradoxes, simplicity, etc. When derivational models permit opaque rule orderings, the mapping between underlying and surface forms is not easily reversible (Cole personal communication). Without an innate derivational grammar, it is difficult to imagine how a learning device could induce an appropriately ordered set of rules from input data. More fundamentally, the postulation of underlying forms presupposes general conditions such as the avoidance of allomorphy, minimization of the number of phonemes, etc.

Finally, the purpose of the syllabification device reduces the impact of the choice of input representations. The syllable network simply syllabifies a string and diagnoses well-formedness, processes that typically apply to the gross structure of the string. Moreover, generative analyses often require syllabification to apply at several different points in the derivation. As a result, the syllabification algorithm must confront a variety of abstract intermediate forms. One possible relationship between traditional rule models and the proposed syllable network is that the former generates and modifies representations while the latter syllabifies them and determines their relative well-formedness.

In addition to coding each of the forms in the corpus into a representational alphabet, we need to determine whether that alphabet comprises primitives or whether they should be further decomposed into features. We have successfully tested a variety of alternatives ranging from treating each segment as an atomic primitive (localist representation), pre-coding each segment with a stipulated featural representation (distributed representation) or allowing the network to induce either some or all of the representation with a number of hidden units. In our earliest work, we decomposed each segment of the inventory into a standard \textit{SPE}-style set of binary features. The sonority vector associated with each segment would be the weighted sum of its active features (see chapter two). Rather than directly modifying the sonority a segment during training, the weight of each of its active features would be modified. While such a procedure permits rapid training of the network, several considerations suggest that the localist or hidden feature alternatives should also be evaluated. Just as the underlying representations in generative grammars often represent theory internal judgments, the featural decomposition of a segmental inventory can also be influenced by a number of theoretical concerns that might the representations unsuitable for the learning network (e.g. separate features for vowels and consonants, binary features selected to provide maximally efficient code, formal vs. observable features, features to simplify rules, capture generalizations or identify natural classes). Most crucially the use of features to define major classes such as syllabic, vocalic, sonorant, continuant, etc., may well prejudice the network since they are
in some sense defined in terms of sonority rank. For instance, a network can learn to
assign a high sonority coefficient to a segment marked +sonorant without claiming to have
accomplished much. Learning that segments marked +syllabic fill the role of syllable
nucleus is nearly vacuous, whether sonority or constituency is used to define the syllable.
To insure that the network's success is not merely the result of the selection of an
idosyncratic or self-validating feature set, we have run simulations using restricted feature
sets that exclude major class features, thereby allowing the distributional work of these
features to be performed by the remaining stipulated features or assigned inductively to a
small set of hidden units.

The fact that the network can appropriately decompose a segmental inventory into
appropriate features serves as an important validation of the autosegmental network. For
the learning of syllabification systems, the featural decomposition is not important,
however. As a result, we will use an atomic (localist) representation for the simulations
described in the remainder of this chapter.

6.1.3 Coding - Output

In addition to selecting an appropriate input representation, we need an
appropriate representation for syllable structure. Since the network architecture proposed
in chapter two includes a column of units for each segment (timing unit), the
representation of syllable structure provides a coding for each segment in the string. For
instance, we could identify onsets (O), nuclei (N) and codas (C), e.g. /goldsmiohl =
ONCCOONC. While such a constituent structure could be mapped from the sonority
wave, the network recognition units, in fact, identify maxima and minima. As a result, we
coda maxima (H), minima (L), and other--either codas or onset clusters (O), e.g. /goldsmiohl
= LHOOLOHL.³

Positing either an ONC or an LOH representation of syllable structures assumes
that the necessary information is independently available to an intelligent system. Since
the target syllable parse is used by supervised training protocols to modify the weights in
the network, the target parse must be available to the system (at least for the training
corpus). It seems quite plausible that the information necessary to identify peaks may well
be independently available to language learners. While the precise cues that are used to
distinguish peaks are unknown, several candidates are available: length, formant structure,
stressability, chest pulses, etc. (Cairns & Feinstein 1982; Maddieson 1985). Young
children appear to have the ability to recognize and mimic the syllable structure of an
utterance absent the ability to produce individual phonetic segments (check references).

³Marking a category "O" (other) that includes codas and non-initial positions in onset clusters may, in
fact, illustrate a genuine linguistic generalization. Vergnaud (xxx) observes that within a given language
the most complex onset typically exceeds the complexity of the most complex codas in monomorphic forms.
The complexity metric may in fact be the number of segments that can be licensed between
maxima and minima on either side of the sonority wave. This would result in the testable prediction that
languages which follow the maximal onset principle (i.e. minimum as initial onset segment), should
permit one more segment in the onset than in the coda (in monomorphic forms).
While information concerning peaks may be primitively available, the information necessary to parse intervocalic clusters has a much more dubious status. As a result, it may be necessary to restrict the representation to H (peak) and O (other). This would make the learning device sensitive to peak information but insensitive to troughs. Fortunately, the learning device can be restricted to the consideration of only peaks without physically recoding the data. As will be noted below, simulations that consider only peak information can successfully model the data. In fact, it is often possible to achieve the parsing of intervocalic clusters without explicit training. For example, the parsing of English intervocalic clusters can be generalized from the distribution of word-initial onsets (seg.ment from *gment).

As discussed in 2.4.3 and 3.4.3, we have also examined the implications of a bipartite syllable representation where we have only the categories left/ascending (U) and right/descending (D), where the initial segment in the onset is (U) and the initial segment in the rhyme (i.e. the nucleus) is (D), e.g. /goldsmed/ = UDDDUUDD. In general, the choice between any of the three above alternatives proves to be descriptively equivalent. Most of our simulations to date have adopted the LOH coding scheme, though recently we have significantly improved the processing speed of the network with equivalent results by adopting the UD scheme (for each learning simulation the syllable coding scheme will be noted). Additionally, the two approaches may make different predictions in dealing with V.V sequences (cf. French 6.4, Spanish 3.4.3). If syllable nuclei are coded as H and the recognition device inherently precludes HH (i.e. two successive peaks without intervening trough), the LOH scheme will necessarily result in errors for the V.V sequences. If, on the other hand, syllable nuclei are coded as D where V.V would be DD, the network respects no similar inherent constraint against DD sequences.

6.2 Learning Algorithm - Early Efforts

While connectionism has provided an extremely powerful arsenal of automatic learning procedures, the phonological phenomena discussed above pose uniquely difficult problems. First, the network deals with complex symbolic objects. Whether a syllabified word is represented as a sonority wave, as an ordered string composed of the atoms L,O,H or U,D or as a hierarchical constituent structure tree (or nested brackets), comparing any two strings, even of the same length, proves difficult. While the network manipulates quantized variables to describe syllables, the difference between two different outputs is not directly quantifiable. Second, even in a network that uses strictly linear activation functions in the input and processing layers, the recognition layer introduces non-linear functions (mini-max, threshold). As a result, a learning algorithm that continuously changes input values will not observe a commensurate improvement in the output. For a given word, the solution space will contain a large plateau representing

4Judgments concerning the parsing of intervocalic clusters are notoriously unreliable even for adult subjects. Dictionaries are similarly inconsistent in assigning intervocalic segments to onsets of codas. As a result, linguistic analyses typically use a theory-internal rationale for making the assignment (e.g. length predictions in Icelandic, Vennemann 1972; aspiration in English, Selkirk 1983).
failure with a smaller region demarcated by steep cliffs representing success. Third, the interactions of the input variables are potentially rather opaque. While increasing the sonority value of a segment definitely increases the likelihood of it becoming a peak, the adjustment of the values of alpha and beta are not nearly so predictable. If the goal is to make a non-peak segment a peak, one could envision several different scenarios whereby alpha and beta could be modified to generate the result.

Taken by itself, the learning of sonority can be relatively simple and straightforward. When a string is processed by the network, the recognition device outputs a string representing its proposed syllabification (e.g. /goldsmiθ/ = LOHLHLHL experimental vs. LHOLOHL target). The learning device processes the string segment-by-segment in order to compare experimental and target parses. It is important to note that the comparison is between abstract objects such as L and H rather than between the sonority values that underlie them, there being no target sonority value for any of the segments in the correct syllabification. As a result, an error term can indicate the direction or sign of the error but not its precise magnitude. Given the possible correspondences below, a coefficient of change could be stipulated for each pair.

(1) Experimental Target Modification (δ)

| H   | H   | .00 |
| * H | O   | -.05 (-.10) |
| * H | L   | -.10 (-.10) |
| * O | H   | +.05 (+.10) |
| O   | O   | .00 |
| O   | L   | -.05 |
| * L | H   | +.10 (+.10) |
| L   | O   | +.05 |
| L   | L   | .00 |

As the learning device processes each correspondence, it can modify the sonority of the associated segment either by an absolute magnitude (ui = u_i ± δ) or by proportion of its current value (ui = u_i ± δ u_i). If the learning device is only sensitive to peaks, only the correspondences marked with asterisks trigger a modification of weights, with the coefficient of change listed in parentheses. With a featural representation, the increment of change is distributed among each of the active features.

In our initial learning simulations, the above procedure was applied to the learning of sonority, but no comparable procedure was available for the learning of alpha and beta. As a result, we stipulated values for alpha and beta and then trained optimal sonority values. Clamping alpha and beta at zero models the ability of the hierarchy in and of itself to account for syllabification. In tests with featural (distributed) representations, the learning of an optimal vector proceeds very quickly (approx. 12-15 training epochs). However, for each of the languages studied (English, Spanish, Berber, Polish), such a model fails to account for a significant percentage of the data (8-30%). As discussed in chapter three, alpha and beta must be permitted to take non-zero values in order to
provide a descriptively adequate account of syllabification. But how are we to train alpha
and beta (and potentially other contextual variables)?

As a brute force testing procedure, we trained the sonority coefficients with alpha
and beta clamped at zero. We then tested the entire corpus of data (400 English words)
on all values of alpha and beta within a given two-dimensional solution space. For
instance, given a precision of .01, we could test all values of alpha and beta from -.30 to
+.30 (3721 value pairs, 1,488,400 total presentations). Given the fact that the sonority
coefficients interact with alpha and beta, the network must iteratively retrain the sonority
vector with alpha and beta clamped at their optimal values. To insure an optimal solution,
the above procedure was repeated three times. For the entire simulation, the network
processed over 4.5 million forms, obtaining a 99.25% performance with $\alpha = -.08$ and $\beta = .04$.

---

Figure 6.2

While the procedure can be significantly streamlined by only repeating the process
twice (no additional performance improvement was obtained on the third pass), and by
reducing the solution space to an "expected" range, the brute force protocol is obviously
implausible as a learning mechanism.
Attempts to stipulate changes in alpha and beta in the same way changes are stipulated for the sonority vector are not nearly as successful, particularly if alpha, beta and the sonority vector are all modified simultaneously. Given the nine correspondences listed in (1), it is not possible to determine which changes in alpha and/or beta are appropriate, either in terms of magnitude or direction. If the window of analysis is enlarged to include more than one segment, we can better predict which errant correspondences could be corrected by given changes in alpha and beta. Unfortunately, if the window includes two segments, we would have stipulate actions for 49 different correspondences; if the window includes three segments, 289 legal correspondences would have to be dealt with. The huge computational cost incurred by such a procedure makes it very implausible. Additionally, the degree of stipulation required runs counter to the spirit of the enterprise. It would ultimately be easier to stipulate the weights in the network than stipulating the procedure that discovers them "inductively."

A rather different and ultimately more successful approach attempts to traverse the alpha/beta solution space intelligently through the use of a gradient descent algorithm (Hinton & Sejnowski 1986; Rumelhart, Hinton & Williams 1986). If the training event is defined as a single presentation of a word to the network, gradient descent will not work because of the plateauing of the energy map. Since the recognition device is non-linear, there is no smooth transition between regions of success and failure. As a result, no matter how large or small one sets the increment of change, no individual change is statistically likely to cross the boundary. Treating an individual form as the training event poses the additional problem that only a small percentage of all of the phonological generalizations necessary for an adequate theory of syllabification are present in any given word.

If, on the other hand, the training event is defined as the presentation of all of the words in the training corpus (essentially viewing the problem as a multiple constraint satisfaction problem with the constraints defined by the correct syllabification of all of the forms simultaneously), gradient descent becomes tractable. At any given point in the alpha/beta solution space, a given number of words will be correctly syllabified. The energy landscape is mapped with alpha and beta as the x- and y- coordinates and the number of correctly syllabified forms as the z-coordinate. Although no smooth transitions can be discerned in the energy landscape when individual words are tested, when success over the entire corpus is plotted, the space can be traversed by an appropriately designed gradient descent algorithm.
Figure 6.3 (contour plot of traversed space)

Figure 6.4 (3-D plot of traversed space)
While this gradient descent procedure possesses several advantages over the stipulative or brute force procedures, several considerations militate against it. Most importantly, considering the training event to be the processing of the entire corpus presents difficulty when applied to natural contexts. While computer simulations can use the entire corpus of data as a single training event, human language learners process words one at a time and appear capable of modifying language hypotheses without going back and testing the hypotheses against every other form in the lexicon.

The next major step in the evolution of learning algorithms is the adoption of stochastic procedures, particularly those that can treat an individual form as a discrete training event, even though it possesses only a fraction of the total information of the network. The graphs in Figure 6.2 and Figures 6.3 and 6.4 both plot success over the entire corpus against the alpha/beta coordinate space. The only difference between the two is the fact that the gradient descent model evaluates only a small subset of the points in the solution space. Might it also be possible to only consider a subset of the training corpus at any or all of the points that are traversed in alpha/beta space? Since individual forms do exemplify a subset of the generalizations found in the total corpus, the contours of the energy space can be plotted dynamically by remembering the success rate for the forms that are processed at any given point while we traverse the alpha/beta space.

To begin the simulation, one begins at a random point in the alpha/beta space. The words are then presented to the network in a random order. If the network syllabifies the form correctly, alpha and beta are left unchanged; if the form is syllabified incorrectly, the values of alpha and beta are changed. In either case, the network "remembers" its results at each point it traverses in the solution space. This can be achieved by constructing a memory buffer that stores the number of presentations and the number of successes at each alpha/beta pair that is tested. The key to the learning dynamic is to use this emerging energy map as a tool to control the modification of alpha and beta. If both the likelihood and magnitude of movement are sensitive to the relative success at the current location, the learning device can gradually converge on an optimal solution (energy minimum/harmony maximum). The following function allows us to model the appropriate sensitivity.

\[
(5) \delta = \text{NORMAL}(\sigma) \ast \kappa \quad 0 \leq \sigma \leq 1
\]

The function NORMAL is a random exponential function that selects a value randomly from a normal frequency distribution with a mean of 0 and a standard deviation determined by the value \(\sigma\). For the proposed learning device, \(\sigma\) represents the failure rate calculated for the current position of alpha and beta. As the failure rate approaches 0, or, in other words, as the success rate approaches 100%, the randomly selected coefficient of change also approaches 0, thereby freezing the network at the optimal global minimum. At points where the network performs poorly, the coefficient of change would be \text{NORMAL}(1), allowing rather large changes. The \(\kappa\) term in the equation is a scaling term. As defined, the normal distribution function takes standard deviation values from 0.0 to 1.0. Since alpha and beta hypothetically range from -0.30 to +0.30, the coefficient of change is clearly too large. The introduction of the scaling term in the equation permits \(\sigma\) to
retain its appropriate 0-1 range while giving δ the appropriate sensitive to train alpha and beta. An additional advantage of the scaling term is the fact that the same basic formula can be used to train the entire range of coefficients that appear in the network. For instance, if sonority coefficients are assigned a range of 0.00-10.00, the same stochastic function can be used for training with simply a different scaling factor. One of the unique features of the proposed learning function is that δ can take both positive and negative values, an appropriate result since we can predict neither the magnitude nor the direction of a change that will produce improved performance. In order to insure that alpha and beta don't take non-convergent values, each must be bounded (-.30 ≤ α, β ≤ .30).

Simulations performed on a 400-word English corpus demonstrate that the stochastic procedure can indeed outperform both the brute force and the gradient descent protocols discussed above:

(6) | Avg. words/ α/β pair | Number of α/β pairs | Total Words |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Brute force:</td>
<td>400</td>
<td>3721</td>
</tr>
<tr>
<td>Gradient descent</td>
<td>400</td>
<td>152</td>
</tr>
<tr>
<td>Stochastic</td>
<td>12.6 (1-632)</td>
<td>1312</td>
</tr>
</tbody>
</table>

In addition to the fact that the stochastic model arrives at a energy minimum much more quickly than the more stipulative alternatives, it allows the learning device to process each presented word as a discreet unit. Despite this fact, the stochastic model described above still poses significant problems both as a learning theory and as a computational model. In order to compute the value of σ for any given point in the alpha/beta space, two values (number of presentations, number of failures) would have to be stored for each point. Since both alpha and beta continuous rather than discrete variables, constrained only by their upper and lower bounds, the memory would have to contain a very large number of cells, limited only by an arbitrarily stipulated precision (cf. gradient descent above).

6.3 Proposed Learning Algorithm - Modified Simulated Annealing

In order to address the difficulties confronted by each of the above approaches, we have developed a unique learning algorithm that successfully models the phonological phenomena under consideration. Stochastic models, such as the one described above, settle into energy minima quite efficiently, but it is difficult to determine whether those minima are local or global. In fact, it is impossible to prove that such a network will find a global minimum (check reference). If the coefficient of change is large, the network may bounce around in the solution space without finding an appropriate energy minimum. If the coefficient of change is small in order to increase its sensitivity, the network may become trapped in a local minimum without sufficient energy to escape. To deal with this problem, simulated annealing procedures have been adopted that dynamically control the likelihood and magnitude of random movement through the solution space (check
reference). In simulated annealing algorithms, the likelihood that a network will move from a particular point in the solution space and the magnitude of that movement are determined not only by the energy/harmony at that point but also by an external variable, temperature. At the beginning of training, the system's temperature is set rather high so that movement through the phase space (solution space) is highly energetic and random. As training continues, the temperature is gradually reduced, reducing the magnitude of movement through the phase space. With the correct cooling schedule, the network will settle appropriately into the global minimum. The principal challenge is to find the correct cooling schedule.

We have developed an adaptation of the simulated annealing procedure that combines the function of $\sigma$ above with a thermostat that modifies temperature dynamically in response to network performance. In our stochastic model, we computed a value $\sigma$ that represented the failure rate at a given point in alpha/beta space. We now replace the $\sigma$ with a temperature variable $\tau$.

\begin{equation}
(7) \quad \delta = \text{NORMAL}(\tau) \ast \kappa \quad 0 \leq \tau \leq 1
\end{equation}

This function can be used to modify each of the various parameters in the network. When used to modify alpha and beta we would have:

\begin{align}
(8) \quad \alpha &= \alpha + \text{NORMAL}(\tau) \ast \kappa \quad 0 \leq \tau \leq 1 \quad -.30 \leq \alpha \leq +.30 \\
(9) \quad \beta &= \beta + \text{NORMAL}(\tau) \ast \kappa \quad 0 \leq \tau \leq 1 \quad -.30 \leq \beta \leq +.30
\end{align}

It should also be noted that the $\delta$ function can be used to modify the sonority vector as well. In the discussion thus far, sonority has been modified according to the procedure outlined in (7) where weights were incremented or decremented by a fixed amount or proportion. The same stochastic procedure can be applied to the learning of sonority, whether the network uses a local (10) or distributed (11) representation for each of the segments in the string.

\begin{align}
(10) \quad u_i &= u_i + \text{NORMAL}(\tau) \ast \kappa \quad 0 \leq \tau \leq 1 \quad 0.00 \leq u_i \leq 10.00 \quad \text{(local)} \\
(11) \quad v_i &= v_i + \text{NORMAL}(\tau) \ast \kappa \quad 0 \leq \tau \leq 1 \quad -1.00 \leq v_i \leq 3.00 \quad \text{(distrib)}
\end{align}

The only difference between (10-11) and (8-9) above is the value of $\kappa$. In each case an appropriate scaling constant must be selected to provide the appropriate sensitivity. In our experiments with syllabification $\kappa = .02$ for the learning of alpha and beta and $\kappa = .125$ for the learning of sonority. While the network will indeed learn the appropriate sonority values with the delta rules listed in (10-11), the speed of learning can be greatly increased by permitting a degree of intelligence into the system. It was noted in 6.2 that unlike alpha and beta, the network can easily determine the appropriate direction of change for the sonority vector though not the magnitude. The formulas in (10-11) randomize both direction and magnitude. If we stipulate the direction of change as in (12) for increased
sonority and (13) for decreased sonority, the learning device will settle into a solution much more quickly.

\[
\begin{align*}
(12) \quad u_i &= u_i + \text{ABS} (\text{NORMAL}(\tau) \times \kappa) \quad 0 \leq \tau \leq 1 \quad 0.00 \leq u_i \leq 10.00 \\
(13) \quad u_i &= u_i - \text{ABS} (\text{NORMAL}(\tau) \times \kappa) \quad 0 \leq \tau \leq 1 \quad 0.00 \leq u_i \leq 10.00 
\end{align*}
\]

Having observed how the temperature variable can be used to determine the magnitude of potential change, we turn to the procedure whereby the temperature in the network is dynamically controlled. Rather than stipulating an externally controlled cooling schedule, the network dynamically changes temperature internally in response to its performance. At the beginning of training, the temperature (\(\tau\)) is set at 1.00. During training, the temperature of the network decays at a constant rate:

\[
(14) \quad \tau = \tau \times \Delta \tau \quad \text{where} \quad \Delta \tau \text{ is a decay coefficient}
\]

For example, if \(\Delta \tau = .99\), the temperature of the network would decay at a constant 1% each time a word is presented to the network. Assuming that this external variable were the only dynamic in the network, \(\delta\) would soon approach 0 as \(\tau\) approaches 0. We propose that temperature serve as an internal dynamic, however, serving as a measure of the marginal rate of success. In addition to being subject to a constant decay (whether a training presentation is a success or a failure), the temperature of the network can be increased whenever a failure is encountered. The same delta rule that modifies the weights in the network can be used to increase the temperature when the network fails.

\[
(15) \quad \tau = \tau + \text{ABS} (\text{NORMAL}(\tau) \times \kappa) \quad 0 \leq \tau \leq 1
\]

Such a formula would determine the potential change in temperature when the system encounters a failure. When the temperature is high, an additional failure significantly increases the temperature and, as a result, significantly increases the potential magnitude of change for future failures. As the temperature declines, a failure not only produces a smaller potential change in the network weights, but also in the amount that temperature could potentially increase. By implication, the temperature variable accurately records the marginal rate of success of the network without recourse to any additional memory buffers.

The precision of the network can be refined even further by making the increase in temperature dependent on the actual change in position in the solution space rather than on the potential change in position. Since NORMAL(\(\tau\)) is a random function and since the change in alpha and beta are independent of each other, the actual magnitude of the movement in 2-dimensional space is determined by:

\[
(16) \quad \sqrt{\Delta \alpha^2 + \Delta \beta^2}
\]
The intuition behind the latter choice of temperature functions is the fact that large movements in phase space should increase the energy dynamic of the system more than small movements. This modification does, in fact, improve the performance of the network, particularly in the vicinity of energy minima (whether local or global).

(17) **Summary of training algorithm**

1. Process word from training corpus using syllable network
   1a. If syllabified correctly, go to step 4
   1b. If syllabified incorrectly --
2. Segment by segment
   2a. If predicted value too low \[ u_i = u_i + \text{ABS}(\text{NORMAL}(\tau) \times .125) \]
   2b. If predicted value too high \[ u_i = u_i - \text{ABS}(\text{NORMAL}(\tau) \times .125) \]
3. Once for word
   3a. Modify \( \alpha \) \[ \alpha = \alpha + \text{NORMAL}(\tau) \times .02 \quad -.30 \leq \alpha \leq +.30 \]
   3b. Modify \( \beta \) \[ \beta = \beta + \text{NORMAL}(\tau) \times .02 \quad -.30 \leq \beta \leq +.30 \]
   3c. Modify \( \tau \) \[ \tau = \tau + \sqrt{\text{SQRT}(\Delta \alpha^2 + \Delta \beta^2)} \quad 0.00 \leq \tau \leq 1.00 \]
4. Decay \( \tau \) by constant rate \[ \tau = \tau \times \Delta \tau \]

As designed, the learning algorithm utilizes a minimum amount of stipulation. The system architecture supplies ranges for each class of coefficients (sonority vector, alpha/beta, metrical inputs, etc.), an associated scaling constant to give the delta rule the appropriate sensitivity, and a temperature decay rate. All other parameters are determined inductively as a result of the training regimen.

6.4 **Results**

To date, the proposed architecture has been applied to seven syllable structure databases: English (400 high-frequency), English (2000 middle-frequency), Berber (81 problem forms), Polish (575 high-frequency children's speech), Spanish (578 high-frequency), Icelandic (65 forms with intervocalic clusters), French (1087 multiple representation).

For the English high-frequency database several tests were run to determine both training performance and generalizability. Two general protocols were followed. First, we tested the network's performance over the entire corpus. Beginning with randomly seeded values for the sonority vector as well as alpha and beta, we trained the network by presenting all of the forms in random order. After each complete presentation (training epoch), the order of words was re-randomized to insure that no ordering effects would

---

*The current implementation of the training algorithm is designed for MS-DOS compatible computers using the SMARTWARE II™ (Informix Corporation, Lenexa, KS) project development language within a spreadsheet environment. Annotated source code can be found in Appendix II. Many of the procedures have also been adapted for use with Borland's Quattro Pro spreadsheet environment.*
appear. The simulation continued until the *temperature* of the network decreased to less than .01. After 100 replications of the learning simulation, we plotted the final values for alpha and beta (Figure 6.18).

![Figure 6.18](image)

During training the system traced the movement of alpha, beta and each of the sonority coefficients through state space, producing a plot of the relative success or *energy landscape* in various regions (Figure 6.19). The results are presented in tabular form below (20):

![Figure 6.19](image)
(20) English (400 words / 100 replications)

<table>
<thead>
<tr>
<th></th>
<th>Avg.</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pct. of correct forms at completion</td>
<td>0.994</td>
<td>0.988</td>
<td>1.000</td>
</tr>
<tr>
<td>Number of training epochs</td>
<td>27.8</td>
<td>10</td>
<td>91</td>
</tr>
<tr>
<td>Value of $\alpha$</td>
<td>0.037</td>
<td>-0.99</td>
<td>0.196</td>
</tr>
<tr>
<td>Value of $\beta$</td>
<td>0.143</td>
<td>0.042</td>
<td>0.247</td>
</tr>
</tbody>
</table>

As indicated, the network rapidly converges on a solution in each of complete replications of the simulation, requiring as few as 10 and no more than 91 presentations of the test corpus to "freeze" in an optimal solution. More importantly, the network's solutions range from a minimum of zero errors out of four hundred words to a maximum of five. It should also be noted that the network correctly parses problem cases such as asked, used, almost, example, six, rest, story, still and against that contain the whole range of possibilities for /s/ and forms like given [gi:v] with syllabic liquids and nasals.

The learning procedure proves to generalizable as well. In addition to running tests where the entire corpus is subjected to training, several split-half tests were run with words randomly assigned to training and test corpora. While only the training corpus was subjected to training, the test corpus was presented at the end of each training epoch for the purpose of comparing relative performance. At the completion of each simulation the network's performance on the test corpus compared favorably to the training set (±2%). Moreover, during training the network's performance over the test corpus generally exceeded its performance on the training set (due to the fact that training performance was measured dynamically so that improvement during a training epoch would not be fully represented). As a final test, we submitted the next 400 words in order of frequency (401-800) and still obtained over 96% performance.

<table>
<thead>
<tr>
<th>(21a) Sonority Rank</th>
<th>Seg. Freq.</th>
<th>Son.</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>29</td>
<td>9.21</td>
</tr>
<tr>
<td>e</td>
<td>33</td>
<td>8.72</td>
</tr>
<tr>
<td>u</td>
<td>30</td>
<td>8.29</td>
</tr>
<tr>
<td>i</td>
<td>72</td>
<td>8.28</td>
</tr>
<tr>
<td>o</td>
<td>48</td>
<td>7.99</td>
</tr>
<tr>
<td>e</td>
<td>78</td>
<td>7.80</td>
</tr>
<tr>
<td>u</td>
<td>43</td>
<td>7.79</td>
</tr>
<tr>
<td>i</td>
<td>44</td>
<td>7.55</td>
</tr>
<tr>
<td>j</td>
<td>67</td>
<td>7.35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(b) Frequency Rank</th>
<th>Seg. Freq.</th>
<th>Son.</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>112</td>
<td>6.01</td>
</tr>
<tr>
<td>n</td>
<td>106</td>
<td>5.31</td>
</tr>
<tr>
<td>t</td>
<td>101</td>
<td>2.60</td>
</tr>
<tr>
<td>s</td>
<td>75</td>
<td>2.48</td>
</tr>
<tr>
<td>i</td>
<td>72</td>
<td>8.28</td>
</tr>
<tr>
<td>d</td>
<td>71</td>
<td>2.66</td>
</tr>
<tr>
<td>l</td>
<td>71</td>
<td>5.96</td>
</tr>
<tr>
<td>j</td>
<td>67</td>
<td>7.35</td>
</tr>
</tbody>
</table>

6Using the present software (see Appendix I for source code) and hardware (486DX2-66) configurations, the network processes approximately 6000 words per minute, meaning that the network converges on a solution within an average of two minutes.
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Frequency</th>
<th>Relative Sonority</th>
</tr>
</thead>
<tbody>
<tr>
<td>æ</td>
<td>32</td>
<td>7.02</td>
</tr>
<tr>
<td>δ</td>
<td>6</td>
<td>6.95</td>
</tr>
<tr>
<td>ρ</td>
<td>20</td>
<td>6.45</td>
</tr>
<tr>
<td>^</td>
<td>13</td>
<td>6.45</td>
</tr>
<tr>
<td>w</td>
<td>31</td>
<td>6.24</td>
</tr>
<tr>
<td>r</td>
<td>112</td>
<td>6.01</td>
</tr>
<tr>
<td>l</td>
<td>71</td>
<td>5.96</td>
</tr>
<tr>
<td>ž</td>
<td>2</td>
<td>5.84</td>
</tr>
<tr>
<td>ŋ</td>
<td>19</td>
<td>5.78</td>
</tr>
<tr>
<td>j</td>
<td>5</td>
<td>5.62</td>
</tr>
<tr>
<td>n</td>
<td>106</td>
<td>5.31</td>
</tr>
<tr>
<td>m</td>
<td>56</td>
<td>4.97</td>
</tr>
<tr>
<td>v</td>
<td>19</td>
<td>4.83</td>
</tr>
<tr>
<td>θ</td>
<td>27</td>
<td>4.78</td>
</tr>
<tr>
<td>f</td>
<td>32</td>
<td>4.72</td>
</tr>
<tr>
<td>g</td>
<td>22</td>
<td>3.88</td>
</tr>
<tr>
<td>h</td>
<td>26</td>
<td>3.45</td>
</tr>
<tr>
<td>k</td>
<td>41</td>
<td>3.42</td>
</tr>
<tr>
<td>ş</td>
<td>5</td>
<td>3.27</td>
</tr>
<tr>
<td>p</td>
<td>26</td>
<td>3.13</td>
</tr>
<tr>
<td>z</td>
<td>33</td>
<td>3.09</td>
</tr>
<tr>
<td>hw</td>
<td>5</td>
<td>2.92</td>
</tr>
<tr>
<td>d</td>
<td>71</td>
<td>2.66</td>
</tr>
<tr>
<td>t</td>
<td>101</td>
<td>2.60</td>
</tr>
<tr>
<td>s</td>
<td>75</td>
<td>2.48</td>
</tr>
<tr>
<td>č</td>
<td>8</td>
<td>2.45</td>
</tr>
<tr>
<td>b</td>
<td>30</td>
<td>0.90</td>
</tr>
<tr>
<td>m</td>
<td>56</td>
<td>4.97</td>
</tr>
<tr>
<td>o</td>
<td>48</td>
<td>7.99</td>
</tr>
<tr>
<td>l</td>
<td>44</td>
<td>7.55</td>
</tr>
<tr>
<td>u</td>
<td>43</td>
<td>7.79</td>
</tr>
<tr>
<td>k</td>
<td>41</td>
<td>3.42</td>
</tr>
<tr>
<td>e</td>
<td>33</td>
<td>8.72</td>
</tr>
<tr>
<td>z</td>
<td>33</td>
<td>3.09</td>
</tr>
<tr>
<td>f</td>
<td>32</td>
<td>4.72</td>
</tr>
<tr>
<td>æ</td>
<td>32</td>
<td>7.02</td>
</tr>
<tr>
<td>w</td>
<td>31</td>
<td>6.24</td>
</tr>
<tr>
<td>u</td>
<td>33</td>
<td>3.09</td>
</tr>
<tr>
<td>w</td>
<td>31</td>
<td>6.24</td>
</tr>
</tbody>
</table>

The relative sonority coefficients that are learned by the network are illustrated in (20a) with the ordered frequency of each segment listed in (20b). It should be noted that both the rank order of segments and their relative sonority values agree very closely with the universal sonority hierarchy described in chapter three (the values in (20a) are the final values from the 100th trial on the 400-word English corpus; if the final values are averaged over all 100 trials, the rank ordering corresponds even more closely to the sonority hierarchy). Moreover, the segments which seem noticeably misplaced (e.g. δ = 6.95; z = 5.84; j = 5.62) are precisely those which have a very low frequency within the 400-word corpus. Since the initial values are randomly seeded with the network only modifying them as much as is necessary to correctly parse each of the forms, high initial values for the low-frequency voiced continuants and affricates would only be reduced a little since they don't appear in clusters. As the corpus grows or as we average several trials, the idiosyncratic values of these segments approach their expected norms.
(22) English (2000 words / 100 replications)

<table>
<thead>
<tr>
<th></th>
<th>.972</th>
<th>.965</th>
<th>.976</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pct. of correct forms at completion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of training epochs</td>
<td>22.1</td>
<td>6</td>
<td>94</td>
</tr>
<tr>
<td>Value of $\alpha$</td>
<td>.017</td>
<td>-.010</td>
<td>.080</td>
</tr>
<tr>
<td>Value of $\beta$</td>
<td>.078</td>
<td>.020</td>
<td>.170</td>
</tr>
</tbody>
</table>

As illustrated in (22), the proposed network proves to have somewhat greater difficulty with the 2000-word medium frequency data base of English than with the 400-word high frequency data base. Rather than being due to a lack of generalizability, the difference is almost entirely due to a single class of forms that appear in the 2000-word list that don't appear in the 400-word list. The former contains a large number of -ly adverbs, including several where the -ly follows a coronal stop (e.g. softly, mostly, frequently, hardly, friendly). Traditional analyses predict that these forms should be syllabified as soft.ly, mos.t.ly, fre.quen.tly, hard.ly and friend.ly because $tl$ and $dl$ can't appear word initially. The phonetic evidence is much less clear, however. In my speech, the network's parses -- soft.ly, mos.t.ly, fre.quen.tly, har.dly, and friend.ly -- more accurately reflect my syllabifications. The assumption that word-initial and word-internal onsets should obey the same constraints may well be an artifact of constituency dependent templatic models (cf. 3.3 discussion of Icelandic where constraints are clearly different; Itô 1986; Larson, *CLS* 1992).

In general, we may need to reassess the correctness of the target syllabifications in those situations where the network is consistently "wrong." The targets are typically the result of some formal theoretical position (Maximum Onset Principle, etc.) rather than clear phonetic evidence. The targets also assume that each form has a single "correct" syllabification. Bernard Laks has compiled a corpus of French forms where two or more potential syllabifications are available for a majority of the forms (find reference). For example, the Laks corpus reports a large number of forms which alternate between word-final consonants and word-final shewa (e.g. aigre [eɡʁ] vs. [eɡ.ʁɛ]) and in the distribution of word-internal clusters (e.g. arbuste [ar.บუ.สํา] vs. [ar.บู้.สํา]).

Given such a corpus, it is possible to test a number of hypotheses. First, we would suspect that if all forms were presented as a training set, the network performance would be compromised by the fact that the network could not simultaneously produce two different syllabifications for a string. We would further expect that the network should be more

---

7In order of difficulty, the twenty-five most problematic forms were: friendly, hardly, frequently, mostly, softly, correctly, directly, rapidly, business, gently, slightly, interesting, newspaper, separate, Atlantic, progress, excitement, established, someone, somewhat, darkness, completely, industry, given, Christmas. A complete list can be found in Appendix II.1.

8In several of the -ly forms, the phonetic realization of the /l/ also appears to be modified, with devoicing not unusual in the /ll/ clusters.

9The network can correctly parse forms with or without final shewa simultaneously, however, because of the fact that the shewa is represented as part of the input. Future research will have to focus on whether
successful on the 519 forms for which there is a single phonetic reflex and syllabification than the 276 forms that have multiple representations. The data in (23a) confirms these predictions. The final column indicates the performance of the network at the completion of training. The fact that the network correctly parses 90.2% of the forms after training is significantly worse than the performance on other data bases (English high frequency - 99.4%; English medium frequency - 97.2%; Spanish - 93.8%; Spanish (excluding VV) - 99.0%; Polish - 96.8%; Polish (excluding intervocalic clusters) - 100%).

The data in the first column illustrates an additional performance measure that allows us to determine the difficulty of learning particular forms. For each word in the training corpus the program records the number of total hits during training. A high value for a form indicates that the network quickly and consistently achieves a correct parse while a low value indicates that the network achieves a correct parse only after significant training or rather inconsistently throughout training. Given the entire French corpus, the network has significantly more difficulty learning the multiple reflex forms (74.8%) than the single reflex forms (81.2%; see Appendix II.3-6 for the performance on each individual form).

(23) French - Multiple Representations

<table>
<thead>
<tr>
<th></th>
<th>Avg. Performance</th>
<th>Avg. Solution</th>
<th>(\alpha/\beta)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) All forms tested (1087)</td>
<td>.780</td>
<td>.902</td>
<td>.10/.15</td>
</tr>
<tr>
<td>Single reflex</td>
<td>.812</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple reflex</td>
<td>.748</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b) Single reflex + alternate #1</td>
<td>.804</td>
<td>.917</td>
<td>.07/.12</td>
</tr>
<tr>
<td>Single reflex</td>
<td>.813</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alternate #1</td>
<td>.770</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(c) Single reflex + alternate #2</td>
<td>.776</td>
<td>.895</td>
<td>.11/.17</td>
</tr>
<tr>
<td>Single reflex</td>
<td>.811</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alternate #2</td>
<td>.708</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(d) Single reflex forms</td>
<td>.810</td>
<td>.923</td>
<td>.08/.17</td>
</tr>
<tr>
<td>(e) Alternate #1</td>
<td>.784</td>
<td>.913</td>
<td>.08/.16</td>
</tr>
<tr>
<td>(f) Alternate #2</td>
<td>.720</td>
<td>.917</td>
<td>.17/.20</td>
</tr>
</tbody>
</table>

The source of complexity in the French data can be further identified by testing various subsets of the corpus. The results in (23b) and (23c) illustrate what happens if a single representation is chosen as the target for each of the forms in the corpus. We segregated the forms with multiple representations into two sets designated as *Alternate #1* and *Alternate #2*. In the first test we combined the set of single reflex forms with the

---

10The assignment of forms simply follows the columnar format found in the Laks corpus (see Appendix II.3). While the columns follow a certain systematicity (e.g. shewa-final forms in the first column vs. consonant-final forms in the second column; onset maximizing forms in first column vs. coda maximizing
Alternate #1 set into a training corpus that was presented to the network using the same training protocol used in (23a). As might be expected, given a single target for each form, the network performed better, both in terms of the final solution (92.7% vs. 90.2%) and in terms of relative ease of learning (80.4% vs. 78.0%). It is also instructive that the network continues to do better on single reflex forms than on multiple reflex forms (81.3% vs. 77.0%), suggesting that if the forms in Alternate #1 were indeed the actual targets their relative difficulty might justify the production of alternative outputs.

When the network is tested on a corpus consisting of the single reflex forms and the Alternate #2 set, a potentially surprising result appears. The network actually performs worse than it does in the original simulation with multiple targets (89.5% vs. 90.2%). The comparison of total hits for single reflex forms (81.1%) with Alternate #2 forms (70.8%) potentially reveals the reason for the relatively poor performance. Apparently the forms in Alternate #2 prove to be more difficult to parse. Since the Alternate #2 set contains a large number of forms with coda clusters, including a significant number that apparently violate the sonority hierarchy (e.g. aigre [egr]), the relatively poor performance is not surprising. As was noted in the discussion of Polish (3.4.2), positive values of alpha and beta permit the network to overcome apparent sonority reversals. The difference between the Alternate #1 test and the Alternate #2 test is found in slightly more positive values for both alpha and beta.

Several observations can be made concerning the above results. While the network gives categorical judgments for each form presented to the network, slight changes in the coefficients or dialectal differences in the input corpus can systematically change those judgments. Alternatively, extraneous inputs (tempo, speech style, etc.) might change the sonority input sufficiently to generate a different form. As a result, the network offers intriguing solutions to speech variability, dialectal differences, historical change and a variety of other phenomena that are less easily captured in more rigid rule formalisms.

6.5 Potential objections

Before concluding the discussion of the learning device, two significant objections must be addressed. On one hand, it might be argued that the network does not, in fact, learn very much. Since the system must have access to a target syllabification for each form in the training corpus in order to identify errors, the network only learns what it, in fact, already knows. Since the language learner must know how a representative sample of words are syllabified before those words can be used to train the network, how can the syllable network be performing any useful work at all when it claims to be able to perform syllabification? This is particularly the case if we presume that the learning device is basically unsupervised. It is not that case that the system produces an experimental forms in second column) there is no implied claim that one column or the other exhibits higher frequency or that one or the other is underlying or more natural. In fact, on theoretical grounds, one might want to correlate consonant-final forms (Alternate #2) with coda maximizing forms (Alternate #1). The simulation simply demonstrates the effects of giving the network a single representation as a target.
sylabification that is evaluated by an external observer that identifies the error (à la Skinnerian conditioning). The system rather has access to an internal representation that has been processed from its environment. The network adds no new knowledge.

This objection can be answered in a variety of ways. First, one could argue that the internal representation available to the unsupervised learner is not the complete sylabification. If the learner only has access to primitive variables such as "long segment" or "vowel" or "clearly defined formant structure" the network can potentially use this information to build a sylabic representation and thereby add information to its primitive representation. In empirical tests, peak-only training proves to quite generalizable to the parsing of non-peaks as well. Second, it could be the argued that the parsed input corpus is a small and simplified subset of the entire language corpus. The fact that the network can generalize from a limited corpus of high-frequency forms indicates that the input does, in fact, possess the structure that is proposed for it. We could go even farther by limiting the input corpus to restricted subsets of the language corpus. If the network could learn to successfully generalize sylabable structure from restricted input samples such as monosyllables, utterance-initial consonantal clusters, utterance-final consonantal clusters, etc., the network would be increasing knowledge by permitting the parsing of any non-restricted input. Third, the primitive information concerning sylabable structure might only be available to the perceptual system, whereas the network provides the mechanism to make this information equally available to the production system. In other words, my ability to successfully identify sylable peaks in utterances to which I am exposed does not necessarily mean that I can sylabify the utterances I intend to produce. The learning device adds to the speaker's knowledge if this interface can be provided. Fourth, even if sylabification in the computational network were to simply recapitulate sylable-structure information which is independently available to the system, the network permits additional phonological processes to be performed and explained (e.g. vowel length, aspiration, licensing, etc.). Finally, the parsed input, even if representative in terms of its composition and informational cues, permits the network to process and produce new forms efficiently. A similar argument appears frequently in the debate between representing a full-entry lexicon and positing generative morphological processes. Even if the latter adds no information, it may be preferable because of its ability to efficiently produce novel forms. Even if children initially memorize all the individual words and even sentences that they understand and use, at some point they develop the generalizations that permit the language user to use both lexical access and generative production of forms. Just as the fact that language users could simply memorize, store, and retrieve all the past tense forms they experience is not an argument against morphology, the argument that language learners know the sylabable structure of a limited corpus should not be an argument against a generalization device such as the dynamic computational network.

A second potential objection to a connectionist learning algorithm is that it can learn too much rather than too little. Critics of connectionism frequently note that an associationist learning device can learn any commensurable set of data whether or not the data represents possible states of the world (i.e. the mind as mush metaphor; cf. Fodor & Pylyshyn, 1988). This criticism is particularly poignant for linguists. A hallmark of the Chomskyan revolution in linguistics is a concern that proposed linguistic theories not have
excessive generative power (Chomsky 1959, 1963). Prince (1992) specifically argues that the proposed dynamic computational network suffers from an inherent overgeneration problem. Considering the model of stress assignment discussed in Goldsmith (1991), Prince observes that the network could describe impossible metrical configurations. For instance, if alpha and beta are equal and both greater than zero, and if the only inputs to the network are a word-initial pulse and a word-final pulse, an arbitrarily long string will always be stressed on the middle syllable (a situation that is unattested and presumably impossible). Moreover, slight changes in alpha and beta would move this single stress leftward or rightward to any arbitrary position in the string.\footnote{Oddly enough, Prince (1992) also credits the dynamic network model with the ability to account for \textit{edge effects}, the ability to restrict stress to a specified window (i.e. the first or last three syllables of the word).}

The discussion of excessive generative power frequently conflates two different but related concerns. First, a \textit{theory} overgenerates \textit{grammars} when it permits grammars that describe impossible languages (i.e. weak generative capacity, Chomsky 1959, 1963). Since an appropriate grammar can presumably not be successfully induced from a finite corpus of data, the theory must formally constrain the set of grammars that can be postulated. Second, a \textit{grammar} for a language overgenerates \textit{structural descriptions} when it permits them to include formulas that are not well-formed (i.e. strong generative capacity). By the latter test, a descriptively adequate grammar is one the accounts for all and only the set of well-formed strings in a language.

While both principles appear frequently in the evaluation of linguistic theories, they should be formally distinguished. Prince's concern is of the former variety. The theoretical formalism (i.e. the dynamic computational network) fails to rule out a grammar that would describe an "impossible" language. Before answering this objection, we must determine whether or not it is an appropriate evaluation metric for the proposed model or even for linguistic theories in general. Strictly speaking, the computational network is neither a theory of language nor a theory of language acquisition. It is simply the device that computes the implications of a variety of theories that could be instantiated in the network (e.g. different activation functions, patterns of connectivity). While theories might be condemned for overgenerating, it is not the fault of the computational device any more than it would be the fault of arithmetic functors that they can potentially describe impossible states of the world. But even if we speak of the formal theory assumed by the use of lateral inhibition to process syllabification and metrical structure or the learning theory assumed in back-propagation or simulated annealing algorithms, we must still ask whether excessive weak generative capacity invalidates the theory.

In the early generative tradition, particularly as exemplified in the \textit{Aspects} model, formal theories of \textit{Universal Grammar} (UG) played a crucial role in constraining both the description and acquisition of grammars. Based on a rejection of both behaviorist and inductive learning theories and armed with the credo that "data necessarily underdetermines theory," generative theory searched for theoretical models that would generate all and only the range of linguistic phenomena that were actually observed in the world's languages and which the linguists' intuitions believed to be possible. Doubting that
data could serve to adequately constrain grammars (especially for language learners) and lacking a learning theory that could itself constrain the range of learnable grammars, generative grammar offered linguists and language learners a highly constrained set of candidate grammars from which to select an appropriate grammar consistent with their limited and imperfect corpus of primary linguistic data.\(^\text{12}\)

Unfortunately, the early concern over excessive weak generative power soon became an article of faith in some quarters. In a defense for its search for "universal phonetics," the introduction to *The Sound Pattern of English* (1968) illustrates the basic argument:

The essential properties of natural language are often referred to as "linguistic universals." . . . The significant linguistic universals are those that must be assumed to be available to the child learning as an a priori, innate endowment. That there must be a rich system of a priori properties—of essential linguistic universals—is fairly obvious from the following empirical observations. Every normal child acquires an extremely intricate and abstract grammar, the properties of which are much underdetermined by the available data. This takes place with great speed, under conditions that are far from ideal, and there is little significant variation among children who may differ greatly in intelligence and experience.

While some version of the above argument prefaces nearly every defense of generative theory, the formal description of the "range of possible languages" (weak generative capacity) was not intended to be the ultimate goal of linguistics, however.\(^\text{13}\) Even in his "Formal Properties of Grammars," Chomsky (1963) writes:

This survey is largely restricted to weak generative capacity of constituent-structure grammars for the simple reason that, with a few exceptions, this is the only area in which substantial results of a mathematical character have been achieved. Ultimately, of course, we are interested in studying the strong generative capacity of empirically validated theories rather than the weak generative capacity of theories which are at best suggestive. . . . In fact, it may well be true that the correct theory of generative grammar will permit generation of a very wide class of languages but only a very

\(^{12}\)Early critics challenged whether the generative formalism in fact provided adequate constraint. Peters & Ritchie (1973), for example, argued that the inventory of available rules permitted generative grammars to describe any recursively enumerable set.

\(^{13}\)In reality, most of the earliest concern over the formal power of linguistic theories concerned the problem of inadequate power rather than that of excessive power. Chomsky (1959, 1963) argues primarily for the generative formalism as a means of increasing the available power vis a vis that available to finite state transducers or phrase structure grammars.
narrow class of systems of structural descriptions, ... Thus the hierarchy of theories that we establish in this chapter (in terms of weak generative capacity) must not be interpreted as providing any serious measure of the richness and complexity of theories of generative grammar that may be proposed (pp. 325-326; emphasis ours).

In more recent work, Chomsky and Lasnik (1991) relax the requirement that a theory of grammar should a priori rule out impossible or non-occurring languages.

... there is no a priori reason to expect that the languages permitted by UG be learnable—that is, attainable under normal circumstances. All that we can expect is that some of them may be; the others will not be found in human societies. If proposals within the principles and parameters approach are close to the mark, then it will follow that languages are in fact learnable, but that is an empirical discovery, and a rather surprising one (p. 4).

In other words, the set of languages that are learnable by the Language Acquisition Device (LAD) only form a subset of those that can be generated by the UG; and rather than then talking about the formal generative power of the LAD, the question of which languages it deems learnable (or learned) is strictly an empirical one. The locus of constraint has moved largely from the formal machinery of a grammatical theory to its attendant learning theory (Wexler & Culicover 1980; Baker & McCarthy 1981; Pinker 1984, 1989; cf. Dresher & Kaye 1990 for specific application to phonology).

While several arguments have been offered for demanding formally constrained theories, the most persuasive has been the argument that formal constraints are necessary precisely because of the inadequacy of empirical constraints. While traditional learning theories can be demonstrated to be inadequate in their ability to induce grammars from evidence, the learning algorithms described in this chapter prove to be very powerful induction devices. If (phonological) grammars of actual languages could be induced from evidence available to the learner, the constraint against impossible languages might also be empirical—the network only learns possible languages because possible (actual) languages comprise the full set of languages that the system encounters. It may well be that other constraints, whether biological, conceptual or historical, prove to be the reason or cause for the distribution of languages that the network is exposed to, but it should not be conceived to be a fault of phonology that its theory doesn't doubly exclude the non-occurring (or impossible) configurations.

Even if formal constraints on the weak generative capacity of a theory were deemed to be necessary, however, the dynamic computational network is not as powerful as Prince's critique suggests. The architecture of the proposed network includes several strong constraints. For instance, the fact that each column of units has an identical pattern of connections and even the same coefficients for lateral inhibition serves as a very strong and somewhat surprising constraint on the kinds of phenomena that can be described. More importantly, a specification of a given network architecture (pattern of connectivity,
number of input units, types of activation functions, etc.) allows an explicit description and measurement of the power of the network, thereby providing the necessary evaluation metric to choose between competing grammars.

As was noted above, learnability is now considered primarily to be an empirical issue. Whether or not the network can or should be faulted for being able to describe impossible languages, it is absolutely necessary to be able to determine whether it adequately describes (and learns) actual languages. As a result, we must evaluate the strong generative capacity of the proposed model. Can the trained network accept all and only the strings that are well-formed within a given language, particularly if it is only trained on positive evidence? Even if the network correctly syllabifies (or assigns metrical structure) to every word that it encounters, does it share the intuition that other strings are not well formed? For instance, the network trained on English must accept the onset cluster /tr/ but be able to reject /tm/ even though it never directly encounters the latter. Otherwise its correct performance on an input corpus may not fully model a speaker's competence when s/he knows which forms are not acceptable in addition to correctly processing those forms that are.

Several lines of response are available. First, in the realm of syllable structure constraints, much if not most of the evidence necessary to exclude impossible clusters is available in the input corpus. For instance, the constraint against the onset cluster *gm is automatically generalized from the correct parsing of the intervocalic cluster in words such as segment. In the discussion in 6.1, we considered the possibility that a network that lacked primitive access to the correct parsing of intervocalic clusters could generalize segment from *gment. If the direction were the reverse, the network would correctly filter ill-formed onset clusters without any direct evidence.

Even if it is not possible to correctly filter "impossible" configurations as a result of experience with related well-formed strings, an additional wealth of negative evidence is actually available to the learning device. If we assume a relatively unsupervised learning algorithm where the network has primitive access to "correct" syllabifications for a finite training corpus, the multitude of errors that the network generates during training serves to exclude a wide range of impossible forms. Generative linguists have generally criticized Skinnerian-type models where an external agent corrects errors generated by a language learner. The proposed model internalizes the process, however, by assuming only that the learner has independent lexical access to a syllabified version of a word that can be compared to the version that is output by the network. Moreover, the comparison need only result in a yea or nay response with the network making a random modification of weights on all errors. If we permit a learning algorithm that makes these assumptions, it would test and reject many more alternative ill-formed hypotheses than any linguist's traditional list of well-formed and related ill-formed strings (designated by asterisks) would likely describe.

It is this latter behavior of the network that offers the most illuminating parallel to practice of contemporary linguistic analysis. One of the strongest justifications for the practice of using speakers' intuitions concerning well-formedness is the fact that they serve as the only reliable barometer of the speaker's linguistic competence (Halle & Bromberger 1991). The argument typically goes as follows: The positive evidence represented by an
actual corpus of data (E-language) fails to fully reflect a speaker's competence, since the speaker has both the ability to both process all of the strings that s/he encounters and pass judgment on strings that s/he would never encounter in a natural context. In order to discover this competence grammar (I-language), the linguist systematically constructs probes of related strings to determine the source of the various grammaticality judgments. Given both the positive and negative evidence, the linguist then constructs a descriptively adequate grammar that generates all and only the set of acceptable strings. Since similar evidence is not available to the language learner, the positive evidence being finite and degraded and the negative evidence non-existent, at least a portion of the resulting grammar must be part of the innate endowment of the child.

The argument is built on two dubious premises, however, particularly with respect to the learning of phonological processes such as syllabification and metrical structure assignment. The critique of induction from a finite corpus of positive evidence relies crucially on the infinitude of the ultimate linguistic domain. While the domain of syntax certainly appears infinite, the realm of syllable structure constraints is very doubtfully so. Even if the sonority hierarchy does not reflect a constraint on the complexity of syllables, they appear to be finite in length, both in terms of competence and performance. While the domains of phrasal intonation or vowel harmony prove to be much longer and not technically bounded, it seems dubious that even these would prove to be infinite. It has been argued, however, that even if unbounded is not equated with infinite we should treat the former as if it were the latter in terms of linguistic competence and computational complexity (cf. Barton, Berwick & Ristad's (1987) critique of KIMMO's morphophonology of Finnish).

Even if we accept the premise that the phonological domain is infinite, the belief that the introduction of negative evidence improves our competence grammar is even more problematic. The typical procedure is to systematically create and test a set of minimally different probes. The operational definition of minimally different is the source of an intractable dilemma, however. It assumes a roughly commensurable number of ill-formed vis a vis well-formed strings.

(23) Taxonomy of possible language types

   a. Finite number of well-formed strings
   b. Recursively enumerable well-formed / recursively enumerable ill-formed
   c. Recursively enumerable well-formed / non-denumerable ill-formed
   d. Non-denumerable well-formed / non-denumerable ill-formed

Of the possibilities listed in (23) we can eliminate (a) and (d) straightforwardly. Given a finite number of well-formed strings, the induction problem doesn't appear. Given a non-denumerable number of well-formed strings, the generative enterprise becomes futile (cf. idealist grammar; check reference). The use of negative evidence assumes that (23b) describes the situation in natural language; an appropriate grammar is not only recursively enumerable, but also recursive (i.e. it is possible in finite time to both accept and reject candidate strings). If (23b) is the case a procedure that selects its probes as in (Figure
8.25) as opposed to (Figure 8.24) appears to be the appropriate way to determine a competence grammar.

![Figure 8.24 and Figure 8.25](image)

Unfortunately, the relationship between the number of well-formed and ill-formed strings may be represented by (23c) instead of (23b). If we assume a grammar containing \( m \) terminals, the total number of strings of length \( n \) would be represented by (26).

\[
(26) \quad \sum_{x=1}^{n} m^x \quad n = \text{length of string}, \quad m = \text{number of terminals}
\]

With a string of unbounded length, if we assume \( m > 1 \), the cardinality of (26) equals or exceeds \( 2^\infty \). In other words, the total number of strings is non-denumerable. As a result, it is impossible to identify a set of minimally different strings for the purpose of eliciting well-formedness judgments, just as there is an infinite set of real numbers between any two rational numbers on the continuum. In fact, the proposed set of ill-formed strings proves to be parasitic on the set of well-formed strings, indicating that a theory of the latter is a prerequisite for even the selection of the former. A learning theory that provides an adequate set of positive exemplars will do as well or better in constraining the set of possible strings as one which randomly selects negative exemplars. The fact that linguists can construct instructive sets of ill-formed probes only proves how well we know the actual grammar of our language.
BIBLIOGRAPHY


Bell, Alan and Joan Hooper, eds. 1978. Syllables and Segments. Amsterdam: North-Holland.


———. 1987. The role of the sonority cycle in core syllabification. ms.


Farmer, Ann K. 1978. The development of a sonority hierarchy: A case for markedness theory. ms. MIT.


Rubach, Jerzy and Geert Booij. Syllable structure assignment in Polish. Phonology Yearbook 7: 118-158.


Touretzky, David S. and Deirdre W. Wheeler. 1989. A connectionist implementation of
cognitive phonology. ms. (paper presented at Berkeley Workshop on Cognitive
Phonology, May 26, 1989).

information processing systems, ed. David Touretzky. San Mateo: Morgan
Kaufmann.

II: Recent progress in connectionist symbol processing. Carnegie Mellon Report
CMU-CS-90-112.


Vaiana Taylor, M. 1974. The great Southern Scots conspiracy: Pattern in the
development of Northern English. In Historical linguistics, vol. 2, eds., J. M.


_______. 1974. Words and syllables in Natural Generative Grammar. Chicago
Linguistic Society Parasession on Natural Phonology, eds. A. Bruck, et. al.
Chicago: University of Chicago Press.

Walli-Sagey, E. 1985. On the representation of complex segments and their formation in
Kinyarwanda. In Studies in Compensatory Lengthening, eds. L. Wetzels and E.

Wexler, Kenneth and Peter W. Culicover. 1980. Formal principles of language
acquisition. Cambridge, MA: MIT Press.

Wheeler, Deirdre W. and David S. Touretzky. 1990. From syllable to stress: A
cognitively plausible model. Chicago Linguistic Society Parasession on the
Syllable in Phonetics and Phonology 26:2, eds. Karen Deaton, Manuela Noske,
and Michael Ziolkowski. Chicago: CLS.

_______. 1992. The role of the syllabifier in Modern Icelandic phonology. Unpublished
ms.


Ph.D. dissertation, University of Chicago.

Mechanical Translation 5:25-41.