Connectionism and Harmony Theory in Linguistics

Alan Prince, Paul Smolensky

CU-CS-533-91  July 1991

University of Colorado at Boulder
DEPARTMENT OF COMPUTER SCIENCE
Connectionism and Harmony Theory in Linguistics

Alan Prince, Paul Smolensky

CU-CS-533-91  July 1991

Department of Computer Science
University of Colorado at Boulder
Campus Box 430
Boulder, Colorado 80309-0430 USA

(303) 492-7514
(303) 492-2844 Fax

Class notes from a 4-week course of the same name taught at the Linguistics Institute, University of California at Santa Cruz, sponsored by the Linguistics Society of America
ANY OPINIONS, FINDINGS, AND CONCLUSIONS OR RECOMMENDATIONS EXPRESSED IN THIS PUBLICATION ARE THOSE OF THE AUTHOR(S) AND DO NOT NECESSARILY REFLECT THE VIEWS OF THE AGENCIES NAMED IN THE ACKNOWLEDGEMENTS SECTION.
Connectionism and Harmony Theory in Linguistics

LSA Linguistic Institute
University of California, Santa Cruz
LINS 247
July 1991

Topics

Spreading Activation & Harmony Maximization

*Spreading activation*

PDP-I:1,2

*Harmony maximization*

PDP-III:3[49-53]; PDP-II:14[7-38]; PDP-I:6

*Goldsmith-Larson Dynamic Linear Model*

Goldsmith, John:

Local modeling in phonology.

A dynamical computational theory of accent systems.

Prince, Alan:


Closed-form solution of the Goldsmith-Larson Dynamic Linear Model of syllable and stress structure and some properties thereof

Inductive Learning & Error Minimization

*Hebb rule*

PDP-I:1[31-40]; PDP-III:4[83-86],5[90-93]

*Delta rule*

PDP-III:[86-89,93-96,121-130]; PDP-I:11

*Past Tense Acquisition Model*

PDP-II:18
Pinker, Steven & Prince, Alan.

Rules and connections in human language.

On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. [Pinker & Mehler.]

*Back Propagation*

PDP-III:5[130-137]; PDP-I:8

*NetTalk*

Sejnowski, Terrence J. & Rosenberg, Charles R. Parallel networks that learn to pronounce English text.

*Recurrent back propagation*

PDP-I:8[354-361]

Elman, Jeffrey. Finding structure in time.

*Continuity & Similarity*

*Hungarian Vowel Harmony Model*

Hare, Mary. The role of similarity in Hungarian vowel harmony: A connectionist account.

*Harmonic Grammar*

Legendre, Geraldine, Miyata, Yoshiro, & Smolensky, Paul:

Can connectionism contribute to syntax? Harmonic Grammar, with an application.

Unifying syntactic and semantic approaches to unaccusativity: A connectionist approach.

Harmonic Grammar -- A formal multi-level connectionist theory of linguistic well-formedness: An application.

Harmonic Grammar -- A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations.

*Harmonic Phonology*

Lakoff, George. A suggestion for a linguistics with connectionist foundations.

Goldsmith, John. Phonology as an intelligent system.

Prince, Alan & Smolensky, Paul. Class Notes

Sources

Unless otherwise indicated, all articles are in the Reader.

‘PDP-X:Y[a-b]’ denotes ‘PDP books, volume X, chapter Y, pages a-b’

On reserve in library & available through bookstore:


Available through Institute:

Reader: Published articles and reports

Class Notes:

To be cited as:

Linguistic Background (incomplete)
LINS 247
Connectionism and Harmony Theory in Linguistics
P. Smolensky/ A. Prince
July, 1991


Kiparsky, P. (ca. 1980) "Vowel Harmony," ms. MIT.


McCarthy, J. and A. Prince (1986) Prosodic Morphology, ms. UMass Amherst and Brandeis University, Waltham.


Optimization in Connectionism & Linguistics

I. Spreading activation & Harmony maximization

II-III. Inductive learning & Error minimization

IV. Continuity & similarity

Harmonic Grammar

V-VIII. Harmonic Phonology

- Universal phonology
  - Typologies
    - Syllable structure
    - Stress systems
  - Phonologies of particular languages

Assignment:
- S-10pp. paper due last class
- Outline due one week before last class

Materials:
- Reader (published papers/reports)
- Class Notes (unpubl. drafts by P&S; handouts; overheads)
- Bookstore
- Library
Summary of 7/8

\[
\begin{array}{c}
1 \rightarrow 2 \leftarrow 3 \\
\downarrow \quad \downarrow \\
1 \quad 2 \\
\end{array}
\]

\[w_{21}, w_{42}\]

- **Nodes**: activation
- **Connections**: weights

\[\text{Net}_i = \sum_j W_{ij} a_j\]

\[a_i = f(\text{Net}_i)\]

**Iteration**

<table>
<thead>
<tr>
<th>Time</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
</tbody>
</table>

**Analogy**: spreading activation \(\sim\) derivation (e.g., autosegmental spreading) \(\rightarrow\) 'structure building'

Goldsmith - Larson D.L.M.

Examples:

\[\alpha \beta\]

\[\text{final } \bar{\theta} = (0 \quad 0 \quad 0 \quad b)\]

- \[1 \quad 0 \quad (b \quad b \quad b \quad b)\]
- \[-1 \quad 0 \quad (-b \quad b \quad b \quad b)\]
- \[-\frac{1}{2} \quad 0 \quad (-\frac{1}{8} b \quad \frac{1}{2} b \quad \frac{1}{2} b \quad b)\]

Prominence
- Sheen (over syllable)
- Sonority (over segment)

Interdigitated nice from nice
Connectionist computation is optimization.

I. The basic (low-level) computational mechanism of conn. is spreading activation.

1. Interesting things happen - in new & different ways - due to this novel mechanism.

   1. Goldsmith & Larson's Dynamic Linear Model (DLM)

   [2... to come]

2. High level understanding of spreading activation - global analysis - is possible through...

   1. Linear Analysis

      1. The activation pattern is a vector in a high-dimensional state space evolving through time - a dynamical system.

      2. The connection strengths form a weight matrix which can be analyzed using linear algebra; e.g., the DLM can be solved in closed form.

      [3. Related methods allow us to use vectors to represent and matrices to process structured data...]

   2. Harmonic Analysis

      1. In certain nets, activation spread steadily increases a numerical well-formedness measure on activation vectors -

         \[ \text{Harmony} = \sum_j \sum_i a_i W_{ij} a_j \]

         \[ \overset{2}{\underset{i} {\text{i}}} \overset{2}{\underset{j} {\text{j}}} \]

      2. Activation spread in these nets is an algorithm for
3. **H** maximization can be interpreted as (soft) constraint satisfaction.

4. These constraints can be used to declaratively define grammar. Global **H** maxima define the grammatical structure (complexity);
   local **H** maxima are computed by connectionist nets (permanence).

3. Activation has several interpretations:
   1. (Graded) truth value of a hypothesis.
   2. Value of a (graded) linguistic quantity
      1. DLM: (continuous) context-dependent prominence

[3. None — meaningless internal computational data]
Harmonic nets:

1. Activation rule: 'If Net > 0, excite, else inhibit'

2. Connectivity:
   - Feedforward
   - Symmetric

3. Updating activations
   - One-at-a-time ('asynchronous')
   - Smooth (unit activations change very little at each step)

Notes:
1. Ensures $\Delta a_i \rightarrow \Delta h_i > 0$ 'each update separately locally bounded'
2. Ensures $\Delta a_i \rightarrow \Delta h > 0$ 'each update separately globally bounded'
3. Ensures $\Delta a_i \rightarrow \Delta h > 0$ 'all connections go up'
4. Ensures $\Delta a_i \rightarrow \Delta h > 0$ 'all connections go both ways'
\[ H = \sum_{i,j} \sum_{w_{ij}} \phi \]

\[ \begin{pmatrix} b_1 \\ b_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ b \end{pmatrix} \]

\[ \begin{pmatrix} a_1 \\ a_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \]

H monotonically increases ... why?

1. Net is feedforward
   'all arrows go up'

2. Only one activation value changes at a time

Constraint satisfaction view:

\[ i \quad \rightarrow \quad \downarrow \]

If it is strong, i should be weak;

" weak " \rightarrow " strong "

Every rule application in the 'direction' is harmonic.
Syllabification in Berber


- onset = 1 segment (except phrase-initially)
- nucleus = 1 segment  NOT NEC. A VOWEL prism: B-r-br
- coda = \leq 1 segment (except phrase finally)

\[
\begin{align*}
\sigma & \sigma \\
\text{tftkt} & \text{yarbbi} & \text{ratki} \\
\text{son} & \text{ons nuc cod}
\end{align*}
\]

What governs syllabification? 'Ideal' sonority profile:

Berber’s 8-level sonority hierarchy:

\[
a > i > \, i > l > m, n > z > f, s, x > b, d, g > p, t, k
\]

\[
8 \geq 7 \geq 6 \geq 5 \geq 4 \geq 3 \geq 2 \geq 1
\]

Dell-El-Medlaoui Algorithm:

For \( s = 8, 7, \ldots, 1 \)

SCAN LEFT-TO-RIGHT

IF FIND \( S_i, S_{in} \) THEN BUILD \( S_i \rightarrow S_{in} \)

ATTACH any remaining unsyllabified segments as CODAS

\[
\begin{align*}
\sigma & \sigma \\
\text{tftkt} & \text{tftkt} & \text{tftkt} & \text{tftkt}
\end{align*}
\]
Harmonic BbrNet

Each $a_i \in [0,1]$ during computation; finally,

- $a_i = 0$ ('non peak')
- $a_i = 1$ ('peak')

$q_i = \pm \{\text{segment} \ i \ \text{is a peak}\}$

Each $b_i$ is a measure of the sonority of segment $i$: (in G-L, $b_i = \text{son}(s_i)$; here, $b_i = 2\cdot \text{son}(s_i) - 1$; $\text{son}(s_i) \in \{1,2\}$, exponential sonority scale in Berber)

Activation rule:

To $a_i$, add $c\cdot\text{Net}_i$; if this is $> 1$, set $a_i = 1$, $c$ small; if this is $< 0$, set $a_i = 0$

Note:
1. ‘If Net $> 0$, excite, else, inhibit’
2. Connections are symmetric
3. Activation updating is smooth

Harmonic

Theorem: 1. For sufficiently large $\lambda$, $\tilde{a}$ is a local $H$ maximum iff $\forall i: a_i \in \{0, 1\}$ and the syllabification satisfies the Berber syllabification constraints: *VVV; *CCC

2. For sufficiently large $\lambda$, $\tilde{a}$ is a global $H$ maximum iff $\forall i: a_i \in \{0, 1\}$ and corresponds to the correct Berber
Harmonic Grammar View of BrbrNet

\[ \cdots \rightarrow -\lambda \rightarrow \cdots \]

Output \( \pm \text{syll} \)

\{ \}

Input \( \pm s \)

\[ \cdot +' = 1 \]
\[ \cdot -' = 0 \]

Grammar of soft rules:

\[ \begin{align*}
\text{a}_i &\rightarrow -\lambda \cdot \text{a}_w \quad \ast \left( +\text{syll}_i \right) \left( +\text{syll}_{i+1} \right) \\
& \quad \rightarrow \text{subtract 'strength' from } H \text{ if violate} \\
& \quad \leftarrow \text{add 'strength' to } H \text{ if satisfy} \\
\text{a}_i &\rightarrow 2^{\text{gon}(s_i)} \sqrt{\left( +s_i \right) \& \left( +\text{syll}_i \right)} \\
& \quad \rightarrow \text{strength: } 2^{\text{gon}(s_i)}
\end{align*} \]

Principle of Harmonic Structural Assignment:

Assign the structure that maximizes \( H \).
Learning from Experience

The problem: suppose we have a collection of input patterns $P_1, \ldots, P_m$, each matched with an output target pattern $T_1, \ldots, T_n$.

Can we get a network to learn to produce $T_i$ given $P_i$?

Concretely, find $w_{ji}$ for

Specifically

We have

$$\text{net}_j = a_1 w_{j1} + a_2 w_{j2} + \cdots + a_n w_{jn}$$

$$= \sum_i a_i w_{ji}$$

And:

- excite out $j$ if $f(\text{net}_j) > 0$
- inhibit out $j$ if $f(\text{net}_j) < 0$

We therefore manipulate $\text{net}_j$ to achieve desired goal.

The inputs are fixed so we work on the weights.

First idea: Hebb Rule:

reward node-wise correlation.

- set $w_{ji} = a_j a_i$
- sum over all input patterns
Technical Background

Vector \( \mathbf{v} = (x_1, x_2, \ldots, x_n) \)

Helps us to speak of chunks

"Pattern" = vector \( \mathbf{p} = (1, 1) \)

\[ \text{dot product} \quad \mathbf{a} \cdot \mathbf{x} = \frac{1}{n} \sum_{i=1}^{n} a_i x_i \]

Example of Hebbian Learning

\[
\begin{array}{c|c|c|c|c}
\hline
p_1 & p_2 & \vdots & \vdots & \mathbf{w}^T \\
\hline
1 & 1 & \vdots & \vdots & \mathbf{w}^T \\
\hline
\end{array}
\]

\[ a = a^+ w \]

\[ \mathbf{W} \]

\[ \mathbf{a} \]
Dot Product measures angle between vectors of unit length.
\[ \vec{a} \cdot \vec{x} = \|a\| \|x\| \cos \theta \]

Perpendicular (Orthogonal) Vectors
\[ \vec{a} \cdot \vec{x} = 0 \]

Dot Product, dividing out length, gives correlation.

Matrix - rectangular array
\[
\begin{pmatrix}
\text{col 1} & \text{col 2} & \text{col 3} \\
\text{row 1} & 11 & 12 & 13 \\
\text{row 2} & 21 & 22 & 23 \\
\text{row 3} & 31 & 32 & 33 \\
\end{pmatrix}
\]

Example
\[
in \rightarrow \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix}
\]

\[
out \downarrow
\]
Hebb en tref.

\[ z(t) = \sum_{t} w_{3i} \cdot x_{i} \]

Two Interesting Cases

\[ (\alpha, \beta, \gamma, \epsilon) = (ax + by + cz) \]

\[ (\alpha, \beta, \gamma, \epsilon) = (ax, by, cy, cz) \]
Multiplying Matrices:

\[ A \cdot B = C \]

The element \( c_{ij} \) of matrix \( C \) is given by:

\[
(c_{ij}) = (a_{ij} + d_{ij})
\]

Adding Matrices:

Rules: Add corresponding elements.

\[
\begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} x & y \\ z & w \end{pmatrix} = \begin{pmatrix} a+x & b+y \\ c+z & d+w \end{pmatrix}
\]
Total Hebb Experience

- Start off: \( w_{ji} = 0 \)
- \( W = \sum_{p} \left[ \sum_{s} \left( \frac{0}{w_{ps}} \right) \cdot (\text{Input}_p - p) \right] \)

Q: What if a trained net faces a new input?

A: Compute

\[
\text{Out}_{\text{new}} = W \cdot \text{Input}_{\text{new}} = \left[ \sum_{s} \left( \frac{0}{w_{ps}} \right) \cdot (\text{Input}_s) \right] \cdot (\text{Input}_{\text{new}})
\]

DOT PRODUCT!

How the trained net behaves:

1. If \( \text{Input}_{\text{new}} \) is all items in training set, output is 0.
2. If \( \text{Input}_{\text{new}} \) is all but one, get a scaled version of the output for that one of the output function, and subtract.
3. If \( \text{Input}_{\text{new}} \) is non-orthogonal, output is a weighted sum of outputs proper to them, weighted by degree of similarity (dot product).

\[
\text{Out}_{\text{new}} = \sum_{k} \left( \frac{0}{w_{pk}} \right) \cdot \text{Input}_k \cdot \text{Input}_{\text{new}}
\]
Using the Delta Rule

\[ W \leftarrow W + \Delta W \]

\[ \Delta W = \sum_{j=1}^{n} \left( e_j \right) \cdot \left( \text{input}_j \right) \]

**Procedure:**
- Present input set \( \{ x_1, \ldots, x_m \} \)
- Output produced by net
- Compare to desired target
- Make step proportional to error \( \epsilon \)
- Do it again
- And again...

\[ a_i = \sum_{j=1}^{n} w_{ij} \]

\[ e_j = \sum_{i=1}^{n} \epsilon_i \]
This procedure will find a set of weights that work, if such a set exists.

\[ E = \frac{1}{2} \sum (y_j - o_j)^2 \]

\[ \Delta w_j = -\eta \frac{\partial E}{\partial w_j} \]

\[ \text{No good} \]

\[ \text{Sign error} \]

\[ \text{Error} \]
Historical remarks

* Learning internal representations - 'hard' vs 'easy' learning
  (Boltzmann learning; Harmony vs 'hidden units')

* Technical breakthrough

\[ \Delta w_{ji} = e \delta_j a_i \]

\[ \delta_j = \text{'error at } j\text{'} \]

\[ = ??? \text{ for 'hidden' unit?} \]

Define recursively: \[ \delta_j = (\sum_k \delta_k w_{kj}) f'(\text{net}_j) \text{ interior} \]

\[ = (t_j - a_j) f'(\text{net}_j) \text{ output} \]

Gradient descent: \[ \Delta w = \text{direction of steepest descent in Error} \]

\[ \Delta w_{ji} = -e \frac{\partial E}{\partial w_{ji}} \]
\[ E = (t-0)^2 \]
Δx ⇔ Δy := a change in x of amount Δx leads to a change in y of amount Δy

\[ \frac{\Delta y}{\Delta x} := \text{rate of change of y w.r.t. x} \]

→ \[ \frac{dy}{dx} \]

for very small Δx, Δy

Chain rule:

Additive property:

\[ \frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \cdot \frac{\partial y}{\partial x} \]

\[ \frac{\partial z}{\partial x} = \sum_k \frac{\partial z}{\partial y_k} \cdot \frac{\partial y_k}{\partial x} \]

Unit k:

\[ \mathbf{a}_k \]

\[ \mathbf{n}_k \]

Unit j:

\[ \mathbf{a}_j \]

\[ \mathbf{n}_j \]

Unit i:

\[ \mathbf{a}_i \]

\[ \mathbf{n}_i \]

Definitions:

\[ a_j = f(n_j) \]

\[ \delta_j = -\frac{\partial E}{\partial n_j} \]

\[ n_j = \sum_i w_{ji} a_i \]
Propagation into the conventional network any:

\[ w_{21} \cdot w_{32} = 1 \]

\[ \mathbf{t}_3 = \mathbf{a}_1 \text{ input:\{0,1\}} \]

\[ \mathbf{a}_3 = w_{32} \mathbf{a}_2 = w_{32} \cdot w_{21} \cdot \mathbf{a}_1 \]

\[ E = \frac{1}{2} \left( 1 - w_{32} \cdot w_{21} \right)^2 + \frac{1}{2} \left( 0 - w_{32} \cdot w_{21} \right)^2 \\
= \frac{1}{2} \left( 1 - w_{32} \cdot w_{21} \right)^2 \]

Start off at \( \mathbf{w} = (0,0) \Rightarrow ??? \)
$z = \frac{1}{2} \left[ 1 - xy \right]^2$
$z = \frac{1}{2} \left( 1 - xy \right)^2$
\[ z = -\frac{1}{2} (1 - xy)^2 \]
\[ z = \frac{1}{2} \left[ 1 - f(y^2f(x))^2 + \frac{1}{2} (y/2)^2 \right] \]
\[ z = \frac{1}{2} \left[ 1 - \left\{ f(y)+f(x) \right\}^2 + \frac{1}{2} \left\{ f(y/2) \right\}^2 \right] \]
Figure 2: Hierarchical clustering results for 48 phonemes.

Figure 4: Fits based on linear regression of the three factor solution on the attribute dimensions "place of articulation" and "vowel height", plotted in Fig. 3. Data based on hidden unit activations. The groupings are based on a centroid hierarchical cluster analysis of the original correlation matrix.
Connectionist Learning

processing architecture +
E function +
minimization technique

\{ \} =
connectionist
learning
algorithm

Extensions of back prop:

- More general processing architectures
- Distal targets
  Jordan: intention (meaning \rightarrow sound) \Rightarrow production (motor) \Rightarrow (sound) \Rightarrow mental model

- Constrained weights
  eg: copies for processing strings
  \[
  \begin{array}{c}
  \text{output} \\
  \uparrow \hspace{1cm} \uparrow \hspace{1cm} \uparrow \\
  \text{input} \\
  \end{array}
  \]
  \[
  \begin{pmatrix}
  w_1 \\
  w_2 \\
  \vdots \\
  w_N \\
  \end{pmatrix}
  \]
  \[
  w_1 = w_2 = \ldots = w_N
  \]

- Recurrent networks

- Different E
  - Occam's razor - simplicity vs accuracy
  - Particular kinds of target patterns
    eg: (0001000) for classifier

Alternatives to supervised learning: no 'targets', new E

- Reinforcement learning
- Unsupervised learning
  - 'Completion task': observe & predict, extracting structure
    No negative evidence

* \rightarrow Elman
Strategies & Principles of Connectionist Applications in Linguistics: A summary

Low-level strategy

1. Connectionist principle:
   Learning is some kind of statistical analysis of co-occurrence.

Linguistic interpretation:
   Temporal sequence of connectionist patterns → temporal sequence of linguistic units

Linguistic principle:
   Linguistic categories can be induced from positive examples thru the prediction of the next item.

Example:
   Inducing word boundaries, lexical categories: Elman

2. Connectionist principle:
   Processing is activation-driven: The [continuous] activation of a unit is increased/decreased over time in proportion to the activation of its neighbors in the network.

Linguistic interpretation:
   activation → feature value (e.g. accentual prominence, sonority, tone)
   network neighbors → tier neighbors

Linguistic principle:
   The [continuous] derived value of a feature is increased/decreased over time in proportion to the derived value of that feature for its left- and right-neighbors in the tier.

Example:
   Goldsmith & Larson theory of stress & sonority

3. Connectionist principle:
   Processing is similarity-driven.

Linguistic interpretation:
   temporal sequences of activity patterns → segmental strings
   connectionist unit → distinctive feature
   similarity → sharing of distinctive features

Linguistic principle:
   The tendency for a missing feature in segment $s$ to be filled with a value consistent with that of a preceding segment $s'$ is increased by
   similarity between $s$ and $s'$
   similarity between $s'$ and other segments near it
   proximity in string/time of $s'$ and $s$
Example:
Transparency in Hungarian vowel harmony: Hare.

Methodological aside:

Low-level strategy for application of connectionism to linguistics (radical):

Low-level connectionist networks [principles] → specific linguistic account/‘model’ [→ higher-level analysis (possibly)]

High-level strategy (conservative):

Low-level connectionist networks → high-level connectionist computational principles → linguistic principles defining a grammatical formalism → specific linguistic accounts

Note: This strategy is made possible only through technical innovations in the high-level analysis of connectionist computation.

High-level strategy

4. Connectionist principle:
Connectionist representations can be analyzed on a higher level as symbolic structures.

Linguistic interpretation:
Henceforth, all grammatical representations will be symbolic: traditional, unrevised. The rules too will be traditional in form, but their interpretation and interaction will be new. (Conservatism.)

5. Connectionist principle:
Processing (at both lower- and higher level) is harmony-driven.

Linguistic interpretation:
connectionist input → linguistic ‘input’ to grammar (string of words, phonemes)
connectionist well-formedness → linguistic well-formedness
pattern of activity arising from an input → structural description of the input
pattern of connections → grammar

Linguistic principle A:
Grammatical well-formedness is quantitatively measured by Harmony. The grammar assigns to any input the structural description that maximizes Harmony. This maximal Harmony value is the well-formedness of the structure: high for well-formed and low for ill-formed.

The terms in the Harmony function define a Harmonic Grammar, and may be interpreted as soft rules of the form:
If structural configuration X occurs in the structure, add $H_X$ to the Harmony of the
structure.
The numbers \( H_x \) can be automatically determined from the data via back propagation.

Note: Harmonic Grammar permits (but does not require) the study of very complex interactions of multiple factors.

Example: Unaccusativity phenomena in French: Legendre, Miyata, Smolensky

Problem: Understanding the grammar is almost as difficult as understanding the original data...

Linguistic principle B:

Optimality. A grammar is a preference relation \( > \) among structural descriptions; it assigns to any input that structure which is optimal (preferred by \( > \) to all other structures).

The preference relation \( > \) among structures is built up compositionally from preference relations among the substructures or subdimensions from which structures are built.

The basic preference relations and their means of combination are drawn from a universal repertoire.

6. Connectionist principle:
Learning is error minimization.

Linguistic interpretation: ???
Example applications: ???
We chose four diagnostic contexts:

| Object Raising (OR) | La glace est facile à faire fondre. | Ice is easy to make melt. |
| cereal | Je croyais Marie déjà sortie. | I believed Marie to have already gone out. |
| Participial Absolute (PA) | Parti avant l'aube, Pierre est arrivé à destination le jour même. | Gone before dawn, Pierre arrived at his destination on the same day. |
| Reduced Relatives (RR) | La neige fondu a formé de la boue. | The melted snow formed mud. |

Specifying HNet' based on linguistic analysis

Data: 760 French sentences generated from the four diagnostic contexts, 143 intransitive verbs, and 3 types of arguments with their acceptabilities judged by human informants in 5 graded levels.

Choice of constituents:
Inputs: We chose the following to code each of the 760 French sentences:
(a) Argument: two semantic features - AN (animate) and VO (volitional).
(b) Verb: two aspectual features - TE (telic) and PR (progressive); 143 individuated verbs as lexicon.
(c) Context: four diagnostic contexts (OR, CR, PA, and RR)
Hidden: two hidden constituents, motivated linguistically by the unergative/unaccusative distinction.

Choice of constituent interactions:
175 independent quadratic terms representing pairwise constituent interactions. (143 lexical, 32 non-lexical)
E.g., we assume each diagnostic context can interact with (have preferences for) semantic features of the argument and aspectual features of the verb (include the terms H=AN, H=TE, etc.).
E.g., we assume each diagnostic context cannot interact with (have preferences for) individual verbs (eliminate terms H=AN, H=TE, etc.)

Treatment of hidden constituents:
We assumed that the two hidden constituents are mutually exclusive: turn on one of the hidden units that maximizes H and turn off the other.

Results

- Examples of soft rules:
  After training the HNet' on these data, we interpret the weights in HNet' as a set of soft rules:
  - verb s'asseoir “walk” prefers hidden constituent 2 - if hidden constituent 1 is chosen, decrease H (well-formedness) by 4.5.
  - verb aller “go” prefers hidden constituent 1 - if hidden constituent 1 is chosen, increase H by 3.4.

- The performance:
  As a collection of these soft rules, the grammar embodied in HNet' could account for the signs [+ or -] of acceptabilities for all of the 760 sentences, and for the graded acceptabilities [5 levels] for all but 30 of them.

- Strong interactions among syntactic and semantic components:
  The network achieved this performance with strong interactions among various syntactic and semantic components; almost every rule is strong enough to flip a decision between acceptable & unacceptableness.

- Extendability of the grammar:
  We tested how well the grammar generated can extend to new verbs. After training on 536 sentences with 87 verbs, we gave the network 80 sentences with 20 new verbs. The network could account for the signs of acceptabilities of all but 2 of the 80 sentences by adding an appropriate new lexical entry for each verb.

- Predictions:
  The network was able to make certain predictions about acceptability patterns allowed for verbs and arguments with certain feature combinations. We are in the process of evaluating these predictions.
Goal: Account for the acceptability judgments of 8393 Ss
(3608 intransive, 4785 transitive)
[an exploration in D.I.C.F.L.; for an exploration in T.I.C.F.L.,
see Prince & Smolensky, 1991]

Data:
- Constructions (11): OR, CR, PA, PE, RR; ON, PCL;
  PRQ, FOR, PR, IMP
- Targets: Subj, D.O.
- Embedded Predicate
  - Identity 408 = 183 transitive + 225 intransitive
  - Features: Given aspectual features - TE, PR
    Learned (arbitrary) features - n = 6
- Embedded Argument Structure
  - Transitives: Subj, D.O., semantic
  - Intransitives: Arg features: VO, AN, DEF
    deep G.R.: Subj, DO.
- Acceptability judgments: +, +?, ?, -, -

Account:
- Implemented as a local connectionist network
  - Trained using backpropagation on judgments
  - Current best account: of 8393 Ss:
    144 errors (1.7%) = 90 transitive (0.5%) + 14 intransitive (+)

Account:
- As a set of Harmonic Grammar rules:
  
  If construction = CR & target = D.O. & deep G.R. = D.O.
  then add 0.5 to Harmony = well-formedness

  If predicate = morir 'die' & deep G.R. = Subj
  then subtract 2.6 from Harmony

  If construction = CR & target = D.O. & Subj = V.O.
  then subtract 1.5 from Harmony

  interacting via the

  Harmonic Structural Assignment Principle:
  assign the structure that maximizes Harmony

- As the higher-level description of a lower-level connectionist network obeying certain general principles of connectionist representation and processing.
  (In particular, pattern completion via Harmony maximization)

Basic linking assumption:
linguistic well-formedness = connectionist well-formedness
= Harmony
OPTIMALITY

Alan Prince
Center for Complex Systems
Boston University

Paul Smolensky
Center for Perceptual Systems
University of Colorado

(1) Optimality in Grammar Choice (Chomsky, 1951, 1965)

UG(D) \rightarrow G_1, G_2, \ldots \quad \text{All } G_i \text{ agree on } D.

Eval (G^1) \rightarrow G^2 \quad G^* \text{ contains the real rules of } L.

Chomsky's idea was to find a coding for the operations and structures of grammar such that the length (i.e., shortest) of the encoding of G was the measure of G's value.

For phonology, we must determine Eval(Lexicon) + Eval(Rules).

Otherwise, we can drive Eval(Rules) to optimum by no Rules.

(2) Optimality as a grammatical mechanism.

Simple Case:

\[ \text{Do } (\text{IN}) \rightarrow \text{OUT}_1, \text{OUT}_2 \quad \text{The real output.} \]

More generally,

\[ B(\text{IN}, \text{Do}a, \text{OUT}^*) > B(\text{IN}, \text{Do}a, \text{OUT}_2) \quad \text{"Greatest Benefit."} \]

(3) The ray against optimization.

a. Have to use numbers — parameter city, and use counting.

b. Many messy tradeoffs if (\# k, \#), what then of ti vs. ka?

c. Optimization is computationally intractable, in general.

(4) Counter-argument.

a. It is not incumbent upon grammar to compute. Well-definition is all; not efficiency or even algorithmicity (Chomsky).

b. Actually ORDER, not counting, is the key.

c. We will argue that many cases of apparent tradeoff can be handled without tradeoff calculation, by a principle of combining orders.

(5) Optimization as a grammatical mechanism: the case of Berber Syllabification (Ddl & El-Medallou, 1987, 1988)

(6) ANY segment may be syllabic:

\[ \text{rAKl}, \ yAKl, tFtKl, AAKk, tZnakk, tZMt, tMzh, tRk, tRk, tRk, tBk. \]

(7) ONSETs: syllables must have onsets, except phrase-initially.

(8) SONORITY: syllabification is sensitive to relative sonority.

a. tZn = \#Zn vs. tMzh = \#mzh

b. ra.sh = \#a.Rzhl vs. luh = \#uLh.

c. D\&E establish the following 8-point hierarchy:

\[ a > u > i > e > r > n > m > z > f > a > b > d > g > p > k \]

low V > high V > liq > nasal > voiced fric > voiceless fric > voiceless stop > voiceless stop

(9) D\&E's Algorithm (DEA)

Form core syllables \((X)_{Y}{\text("CV")}, \) where \(X\) is any segment and \(Y \) is successively replaced by the features describing the steps of the sonority scale; in descending order.

Iterate from Left to Right for each fixing of the nuclear variable \(Y\).

(10) Secondary phenomena (to be ignored):

a. obstruents de-syllabilize \# and \#.

b. sonorants \(C\) optionally desyllabify \#.

c. automorphic geminates never begin an onset. (\# + m)

(11) EXAMPLES:

\[ /l\text{t\text{\$}}\text{\$}/ \rightarrow [\text{la}]t\text{\$} \rightarrow [\text{t}]\text{\$}\text{\$} \rightarrow [\text{t}]\text{\$}\text{\$} \rightarrow \text{"thou stored"} \]

\[ /\text{t\text{\$}}\text{\$}\text{\$}/ \rightarrow [\text{t}]\text{\$} \rightarrow [\text{t}]\text{\$}\text{\$} \rightarrow \text{"thou stored"} \]

(12) DEA serves to optimize the quality of syllables: high sonority items like to be nuclei.

(13) But DEA per se makes no contact with this. Eg. Why descend the scale?

(14) Harmonic Syllabification (HS).

Form the best core syllable from free material in the domain, good \((\#) > \) good \((\#)\).

Iterate until you can do more, \[ H_{\text{S}}(C) = E. \]

(15) This is a fully 'Harmonic' process: 'Harmony' from Smolensky, 1986."

Go to the most well-formed available state.

(16) Note that HS is different from DEA. HS has no directional bias, does not sensibly run in a direction (LR) to deal with sonority plateaus.

(17) Evidence for directionality weak anyway.

a. \(/\text{k\text{i\text{u}}}/ \rightarrow \text{ki}, \text{kyu} \quad \text{but better onset (DE note)}

b. \(/\text{\text{\$}}\text{\text{\$}}/ \rightarrow \text{\$\text{\text{\$}}}, \text{\$\text{\text{\$}}} \quad \text{but no final syllabic obstruent.}

c. \(/\text{\text{\$}}\text{\text{\$}}/ \rightarrow \text{\$\text{\text{\$}}}, \text{\$\text{\text{\$}}} \quad \text{but no geminate qua CV.}

d. \(/\text{\text{\$}}\text{\text{\$}}/ \rightarrow \text{\$\text{\text{\$}}}, \text{\$\text{\text{\$}}} \quad \text{both ok, DEA prefers only m, \#d.}

e. \(/\text{\text{\$}}\text{\text{\$}}/ \rightarrow \text{\$\text{\text{\$}}}, \text{\$\text{\text{\$}}} \quad \text{DEA fails. Nb. closed \# bad.}

(18) Conclusion: optimization, with grades of well-formedness (representational Harmony), is a mechanism of grammar.

(19) To make this approach fly, we need two things:

a. Ways of evaluating combinations of phonological dimensions where the dimensions themselves needn't have an internal well-formed structure, but may be scalar. Harmony scale on dim. comb. ([post-Praguean context markedness: SPE, Kean, ...Archeangeli & Pulleyblank]

b. Ways of combining different relevant dimensional scales to one metric scale. (eg. [O'Connor] 7 + 10 = 17, where 7 = Harmony, 10 = Stress, 17 = HARMONY + STRESS).


a. Separate prominence (not markedness) scales: \(H \geq L\), stress \\( \succ \) unstress.

b. Combine Implicatively: if \(H\), then stressed. (\(\text{WSP} = \text{Hayes's 'quantity sensitivity'}.\))

Good are: \(H^*\), \(L\). Bad is: \(L^*\). (unstressed light)

c. Weaker implication: if stressed, then Heavy.

Good are: \(H^*, L^*\). Bad in: \(L^*\). (unstressed heavy)

d. Combination respecting strengths of implications: \(H^*\) \(>\) \(L^*\) \(>\) \(H\). This is a Harmonic ordering, expressing degree of coarseness of the two dimensions.
(21) Combination of Harmony Scales. Example: 'unbounded' stress systems like: Stress the rightmost heavy syllable, if no heavy then stress the rightmost syllable (R/R). Prince (1983) and Hayes (1998) see these as enhancements of prominence rather than 'unbounded foot' phenomena. We extend this result, thereby limiting feet to the authentic binary rhythmic units.

(22) The pattern R/R. Two factors:
  a. WEIGHT. Heavy syllables should be stressed.
  b. POSITION. Edgelessness: Stress should be near the end.

(23) WEIGHT dominates POSITION. WEIGHT >> POSITION. (N.b. One stress.)

(24) WEIGHT must be satisfied, if possible. But there can be several choices:
  a. cv.CVv.cv.cv.cv.cv.
  b. cv.cv.cv.cv.CVv.cv.cv.

(25) Indeterminacies are resolved then by POSITION. (24b) is chosen. 4b or 4b.

(26) No incertimeraency in (cv) = ...cv.cv.cv.

(27) This is LEXICOGRAPHIC ORDER — a generalization of alphabetic order. Alphabetically, strings are ordered first by initial character (weight), then ties are broken by the ordering on the second character (weight), etc. Thus 1 = <a=b>c. NE-NO TRADEGEP.

(28) LEXICOGRAPHIC PRODUCT of two orders, F * >> O, total order on the the pairs (pa) constructed from the orders on P and on Q; first ordering by F-dimension, then by Q-dimension.

(29) Example: alphabetic ordering itself: POSN * >> ALPHA.
(1,2,3,4,...) >> (a,b,c,...)
1 > b > 1 > c >> 2 > c.

(30) Claim: Lexicographic Order is the essential mode of combining Harmony scales.

(31) Berber Syllabification is also a kind of Lexicographic Order: analyses in which /m/s are nuclei are absolutely preferred to competitors in which /v/s are not but /v/s are.

(32) Prominance-driven Stress. Claim: works by Lexicographic ordering PROMINENCE >> POSITION, with development of notions of prominence and position.

(33) Hindi (modern) (Kelkar, 1968; Hayes, 1991)
'Stress falls on the heaviest available syllable, and in the event of a tie, the rightmost nonfinal candidate wins.' (Hayes)

(34) Heaviness: CvC, CvCC (μ) >> Cv, CvC (μ) >> Cv (μ)

(35) Heaviness is best:
\[ \text{kiDIAH} \quad \text{jaKAA} \quad \text{as.BAA} \quad \text{ru.pi.AA} \quad \text{REEZ.gaa.nii} \]

(36) Among equals, last nonfinal is best:
\[ \mu \] ru.KAA.yaa FUS.ta.kee roo.ZAA.naa
\[ \mu \mu \] AA3.maa.jaah aas.MAA.AN.jaah

(37) Analyze: WEIGHT >> POSITION
But POSITION is constructed from NONFINALITY, and RTMOSTNESS.

(38) NONFINALITY >> RTMOSTNESS

(39) WEIGHT >> NONFINALITY >> RTMOSTNESS.

(40) Piraha. (Everett & Everett,84; Everett,88; Levin, 85; Davis, 88; Hammond, 90; Hayes, 91).

"Stress the rightmost token of the heaviest syllable type in the last 3 syllables of the word."

(41) Parallel: "rightmost heavy or rightmost syllable". 'HEAVINESS' >> EDGEGMOPHT.

(42) What is 'Heaviest'? C = voiceless, G = voiced cons.
a. ?A ba.gi 'touscan' CV+Gv
b. ?A ba.PA 'Amapa' rnot+Gw
b. sa.Al.pi 'josp' cv w >> Cv
c. KAA.gai 'word' Cv+Gv
d. pia.GAI hi.ni 'banana' Gw+Vw ; in 3σ window; nb poo.
e. ka.pi.gi ka ga.kai ka ba.BA 'ritmost in window: std. kai
f. ka.pi.gi ka ga.ka.gai ka ba.BA 'ritmost in window: std. kai

(43) Full 'Heaviness' Hierarchy:
Cv v >> CV v >> CV >> Cv >> V , where C=voiceless, G=voiced.

(44) Evidence that the stress is these:
a. Some speakers treat it as a beat wth hand-gestures.
b. Some speakers devote everything after stressed syllable.
c. Some speakers delete everything after stressed syllable.
d. May be some tropism of tone to stress.

(45) Issues:
A) How is 3 syllable limitation established?
B) How is 'heaviness' hierarchy established?

(46) Minimal about 3? Prosodic Hierarchy says Wd contains F contains σ. So, hierarchically-locally, minWd contains ONE F. Maximal such Wd = F + σ (or indeed σ + F).

(47) Piraha stress must lie in maximal minWd, for F = σ.

(48) Evidence for involvement of Wd category: Closely-knit phrases show ONLY ONE STRESS when last word is 2 σ.
ba.sa.i # bi.SAI 'cloth # red'
\(7a.ba.pa \# \text{go.GI} \) 'city-same # where-at'
\(7l.sii.HOA. \# \text{ai.BAI.7i} \) 'liquid fuel # much'
\(7a.pa.PAI \# \text{H1a.ba} \) 'head # hurts'
\(ka.H(AJ) \# ?o.ga.BA.GA \) 'arrow # want'

Analysis: FWd = max minWd is established at -edge of morph-word.

3

4
(49) PWJ Parsed. Stress falls inside PWd. Must have #.

b.o. [sai hi SAi] #
[ts pa.PA] #
[tl.ia ba] #

(50) ANS. to (48). 'Heaviness' Hierarchy
CVV > GV > CV > GV > V
a. CVV > GV in every case.
b. C > G > mil This is just onset goodness (R. Slevin).

Prominence Order = Weight >> Onset

(51) Full Story. IN-PWD >> WEIGHT >> ONSET >> RIGHTMOSTNESS.

Stress penultimate, final, antepenultimate, dependent on foot structure AND VV location.

In words with only SHORT VOWELS, stress is penultimate:

Sata pa.Tlta dhana.HEra bhi.na.Sera
jaa.han.NEma sun.HARA

This is actually a Moraic-Trochee (RL) pattern (Hayes) - weak vowels in open syllables reduce; those in closed do not; nb. no closed final syllables.

(53) LONG VOWEL FINAL: Stress is FINAL. For SHORT penult V-

a.dina.TAA gora.min.Tli eka.bat.TAA

LONG V in antepenultimate stressed if VV followed by two LIGHT:

ji.muta.BAA.baa kaka.ROO.hari BAA.sana

PENULT LONG V is stressed:

pu.RUU khA SAA.bhaa ba.rha.MAA.saa

(54) ANALYSIS: Main Stress falls on MOST PROMINENT in final circumflexed window; when there is equality, penult is favored (Non-finality).

(55) Window: 2 units = FF, Pa, or F, = 2 F, if headless feet exist (HB6)

(56) Elementary Prominence Orders:
FOOT: Foot-heads > Nonheads;
VV: VV > V

Relation: VV >> FOOT

(57) As usual, PROMINENCE >> POSN.

(58) For POSN = NONFINALITY >> RIGHTMOSTNESS

(59) Cases:

[CVV.CVv] = prom/rightmost Fhead. dhana.HEra
di. to ja.ha.NEma
cw.CV.CVv] prom wins. ku(u).ra.NAA
[JCVV.CVv] prom wins.

(60) Fact Note: Following Hayes, we assume [CVV.CVv]
gala + bennA -> galaBenAA but CAA.rilo 'fourth' VV shortens in nonfinal, non-mainstressed syllables. If wrong, then PROM = heavy.


(62) Broad Vista. To run grammar on optimization, we need to be able to say WHEN to perform an operation. I.e. when it is better to act than to do nothing:
B (X -> Y) > B (X -> X)


(64) From universal perspective, need to order Transitions and combinations of Transitions and contexts to obtain markedness implications about processes.

(65) Example. LANGO Vowel Harmony = one property thereof. (Pulleyblank, Yeti)

a. ATR Harmony spreads forward from ALL vowels over a SINGLE CONSONANT: in V1,C-V2 , V2 harmonizes with 1st vowel.

b. But over TWO intervening C's, he trigger V must be high:
only in hi-C-CV do we get spread of ATR.

(66) Appears to involve complex TRADE-OFF in virtues of spreading from various sources, and difficulty of doing same in various environments. But there is NO TRADE-OFF.

(67) Approach. Suppose we know the following:

a Spread ATR from HI > Spread ATR from Non-HI. "<X M X >" b. Spread over single C > Spread over CC. = Distance"

(68) Form the Lexicographic Product

Spread from HI over C > Spread from H over CC

= Spread from Non-HI over C > Spread from Non-HI over CC.

(69) Now, an individual language may insert DO-NOTHING anywhere in the order. All things preferable will be done, then.

(70) Insert DO-NOTHING at bottom: any action is better than inaction: no constraint, full ATR harmony.

(71) Lango. Insert Do-Nothing above last element in order. Spreading-from-Non-HI-over-CC is worse than Nothing: But everything else is better. So ac. accordingly. ***
Constraint Interaction and Harmony Maximization in Grammar

Alas Prince
Braadel University

Paul Smolensky
University of Colorado

I. Basic Idea -- General
A. Picks up two themes in modern linguistic thought (see attached bibliography; most of the entries are concerned with these):

1. Constraints or processes that give way or are conditional.
   a. Do such-and-such EXCEPT when the result conflicts with some other constraint. Eg. Make final or extrametrical EXCEPT in monosyllables. Eg. Foot form.
   b. Do such-and-such ONLY when necessary to achieve wellformedness. Eg. Epenthesis a glottal stop ONLY to provide syllable with onset. Epenthesis a vowel ONLY to make word big enough to hold foot.

2. Direct application of universal principles of markedness in the operation of individual grammars.
   a. Prefer syllables to have onsets. Prefer the collocation of "HI and ATR. In the standard theory (eg. SPE, ch. 1-6, et seq.), there are no more than observations made about what grammatical processes happen to end up doing. But they can form the foundation for principled theory of grammar itself; can be directly appealed to in specific grammars.

B. Rapport with model of functioning of central species of connectionist networks: harmony increase to maximum.

C. Fundamental Harmonic Imperative: among the possibilities, choose the one that maximizes harmony (relative wf'edness).

II. What we need to carry out this project.

1. Notion of preferences (approx = markedness)
2. Ways of combining orders. (Lexico. Direct.)
3. Way of assessing harmony of structures, as determined by constraint order. (LOCSS. Tableaux)

III. Constraint Domination. Some familiar basic types of constraint interactions.
A. Do something EXCEPT. A positively-regarded option is embedded lowdown.

1. Extrametricality: mäns, *(män)> LexWD=PrWD >> EXM.

LexWD=PrWD. Lexical Words should be Prosodic Words, therefore contain feet.

EXM. Final syllables should be foot-loose.

2. Imbrication in Floma (Hyman, 1991):
   Short Base Long Base
   fik-ll -e *matik-ll-e
   *fīl-1-t -e matil-1-k -e

Generalization. Stems CvC suffix ll; Longer bases insert l after last vowel in bases.
Analysis. Take l to be default C. Then suffix /l/ to PrWD.

FB >> PrWDWF

FB. Feet should be binary at some level of analysis (go, Mh).
PrWDWF. PrWD words should be consist exhaustively of well-formed units.

3. Foot WFedness. (fél, *fél, *fél)

FB >> LexWD=PrWD >> EXM >> PARSE >> RH

RH. Rhythmic Harmony. Trochodial feet LL, H are better than LH.
PARSE. All material should be parsed into prosodic structure. (= Itô's Prosodic Licensing).

Say Stray Erasure en route to the Phonetics.

   Goodfirstyll >> Leftmostness of prefix

Goodfirstyll. No Coca.
Leftmostness of Prefix. 'Prefix' means affixal content is Leftmost in Affix + Stem collocation.

   Edge-preserve >> Rightmostness of suffix (light syll).

Edge-preserve. Segment at edge of base should appear at edge of reduplicated base (pa-prath)
Rightmostness of Suffix. 'Suffix' means affixal content is Rightmost in Affix + Stem collocation.

B. Do something ONLY. A rejectionist option is lowdown & overruled by a higher-order opt - that is, the high-order constraint can force the disfavored option.

HAVE-ONSET >> FILL-NODES

HAVE-ONSET = a should dominate ON constituent.
'Epenthesis' = Insert empty syllable structure (Selkirk, Itô).
FILL-NODES = nodes should dominate segments - AVOID epenthesis.
Universal Typology in Harmonic Phonology: 
0th-order Syllable Structure

Paul Smolensky & Alan Prince
July 25, 1991

Fundamental Definitions

LOCS:
Suppose given:

- a role-filler decomposition of \( S \) specifying for each \( s \in S \) a description of \( s \) as a multi-set \( M_s \) of role/filler pairs
- an ordering \( \succ \) of fillers
- a function \( \text{predom} \) specifying a predominant element of any multi-set \( M_s \) of role/filler pairs; define:
  \[
  \text{residue}(M) = M \setminus \text{predom}(M)
  \]

Then the corresponding Lexical Order on Composite Structures \( S \) is the ordering on \( S \) defined recursively by:

\[
M_i \succ M_j \iff \begin{align*}
\text{filler}(\text{predom}(M_i)) &= \text{filler}(\text{predom}(M_j)) \\
\text{or} \\
\text{filler}(\text{predom}(M_i)) &= \text{filler}(\text{predom}(M_j)) \text{ and } \text{residue}(M_i) \succ \text{residue}(M_j)
\end{align*}
\]

Two cases of interest:
- Filler-driven LOCS: \( \text{predom}(M) \) = any role/filler pair in \( M \), containing a maximal filler (w.r.t. \( \succ \)).
- Role-driven LOCS: \( \text{predom}(M) \) = a role/filler pair in \( M \), containing a maximal role, w.r.t. an ordering on roles \( \supseteq \) which must be further supplied

SDG: The Strict Domination Grammar

\( p_1 \succ p_2 \succ \ldots \succ p_n \)

is the preference relation \( \succ \) defined from a sequence \( p_1, \ldots, p_n \) of Boolean predicates over possible structural descriptions, defined as the following role-driven LOCS:

- the ‘roles’ are the predicates \( p_1, \ldots, p_n \) and the ‘filler’ of each role is the truth value of each predicate: \( M_s = \{ \pm p_1, \pm p_2, \ldots \} \) where the sign of \( p_k \) indicates whether \( s \) satisfies predicate \( p_k (+) \) or not \((-). \) [We sometimes write \( M_s \) more mnemonically by suppressing the \( +p_k \) and writing the \( \neg p_k \) as \( \ast p_k \)]
- the ordering on fillers is \( + \equiv (T \equiv F'; \text{‘preferences’}) \)
- the ordering on roles is \( p_1 \supseteq \ldots \supseteq p_n \)

[Aside: If in the definition of SDG we change the LOCS from role- to filler-driven, then the order we get is ‘Prefer fewer marks’, where a ‘mark’ is a violated preference.]

FUT(\( P \)): The Factorial Universal Typology determined by a set \( P \) of universal preferences is the typology arising from the strict domination grammars determined all possible permutations of the preferences in \( P \).

[That is, as far as the grammatical module governed by the set \( P \) is concerned, FUT(\( P \)) characterizes the set of possible languages as exactly those arising from all possible orderings of the preferences in \( P \).]
0th-order (C/V) Syllable Structure

Goal:
To explain the typology:
- onsets: required/optional/ * forbidden;
- codas: * required/optional/forbidden

[nuclei: assume required by definition of syllable]

<table>
<thead>
<tr>
<th>onsets</th>
<th>required</th>
<th>optional</th>
</tr>
</thead>
<tbody>
<tr>
<td>codas</td>
<td>forbidden</td>
<td>$\Sigma_{CV}$</td>
</tr>
<tr>
<td></td>
<td>optional</td>
<td>$\Sigma_{CV(c)}$</td>
</tr>
</tbody>
</table>

Main Claim:
The universal typology of C/V syllable structures is \( \text{FUT}(P) \), where the set of preferences \( P \) is:
- ONS: Prefer syllables with onsets.
- -COD: Prefer syllables with no coda.
- PARSE: Prefer parsed segments.
- FILL: Prefer filled syllable positions.

Note: Best syllable is CV, which satisfies all preferences. Thus /CV/ is assigned CV. So it is not possible to prohibit onsets or to require codas.

Onsets (ignore -COD for now):
Look at underlying form with preferred coda (i.e., none) and an onset wanting repair (i.e., none): /V/;
Possible structures:

```
  \sigma
    \downarrow
  NUC
    \downarrow
  \uparrow
  V.
    \{ * ONS \}
```

```
  \sigma
    \downarrow
  ONS
    \downarrow
  NUC
    \downarrow
  \uparrow
  V.
    \{ * PARSE \}
```

```
  \sigma
    \downarrow
  \uparrow
  \^V.
    \{ * FILL \}
```

Phonetic assumptions: unparsed segments are not realized (deletion); unfilled syllable positions are somehow realized (epenthesis).

The preferred structure is the one with the lowest mark; given an order
\( p_1 \gg p_2 \gg p_3 \),
the preferred structure is the one having \{ * \( p_3 \) \}; the order of \( p_1 \) & \( p_2 \) don’t matter. So rather than 3! = 6 cases there are 3 different ones (ignore \{ \} for now):

1a. ONS \gg \{ PARSE \gg FILL \} \sim PARSE \gg ONS \gg FILL  
    Preferred: \^V.

1b. ONS \gg \{ FILL \gg PARSE \} \sim FILL \gg ONS \gg PARSE  
    Preferred: [V]

2a. \{ PARSE \gg FILL \} \gg ONS \sim \{ FILL \gg PARSE \} \gg ONS  
    Preferred: V.

(Here, \sim \) denotes the equivalence of two orderings, that is, that they determine the same maximally harmonious forms.)
Using the first instance of each case, we can view the three cases hierarchically by letting

1-1: Prefer one-to-one pairing of segments and syllable positions

= either \{\text{PARSE} \gg \text{FILL}\} or \{\text{FILL} \gg \text{PARSE}\}

we get:

1. ONS \gg 1-1
   a. 1-1 = \{\text{PARSE} \gg \text{FILL}\} Preferred: ^V.
   b. 1-1 = \{\text{FILL} \gg \text{PARSE}\} Preferred: [V]

2. 1-1 \gg ONS
   1-1 = \{\text{PARSE} \gg \text{FILL}\} - 1-1 = \{\text{FILL} \gg \text{PARSE}\} Preferred: V.

That is:

1. Onsets are mandatory, since ONS is the predominant preference
   a. an onset is epenthesized;
   b. an onsetless nucleus is deleted.

2. Onsets are optional, since it is more important not to epenthesize or delete (1-1) than it is to have an onset.

Codas:

Same story, \textit{mutatis mutandis}. Look at /CVC/ (preferred onset, coda needing repair).
Possible structures:

\[ \sigma \]
\[ \text{ONS} \quad \text{NUC} \quad \text{COD} \]
\[ c \quad v \quad c \]
\{ = -\text{COD}\}

\[ \sigma \]
\[ \text{ONS} \quad \text{NUC} \]
\[ c \quad v \quad c \]
\{ = \text{PARSE}\}

\[ \sigma \]
\[ \text{ONS} \quad \text{NUC} \]
\[ c \quad v \quad c \]
\{ = \text{FILL}\}

Cases:

1. -COD \gg 1-1
   a. 1-1 = \{\text{PARSE} \gg \text{FILL}\} Preferred: CV.C^.
   b. 1-1 = \{\text{FILL} \gg \text{PARSE}\} Preferred: CV.[C]

2. 1-1 \gg -COD
   1-1 = \{\text{PARSE} \gg \text{FILL}\} - 1-1 = \{\text{FILL} \gg \text{PARSE}\} Preferred: CVC.

That is:

1. Codas are prohibited, since -COD is the predominant preference
   a. an extra nucleus is epenthesized;
   b. a coda is deleted.

2. Codas are optional, since it is more important not to epenthesize or delete (1-1) than it is to lack a coda.

Note: For the onset discussion, the relevant case of FILL is 'Prefer filled onsets' (\text{FILL}_{\text{ONS}}), while for the coda discussion, it is 'Prefer filled nuclei' (\text{FILL}_{\text{NUC}}). These need not be assumed to be subsumed under a single general preference, FILL. Similarly, the relevant cases of PARSE are \text{PARSE}_{\text{v}} and \text{PARSE}_{\text{c}} for the onset and coda discussions, respectively. If there is but one general preference FILL and another such PARSE, then these general preferences must be relatively ranked once and for all in a given grammar, predicting that, \text{PARSE}_{\text{v}} \gg \text{FILL}_{\text{ONS}} iff \text{PARSE}_{\text{c}} \gg \text{FILL}_{\text{NUC}}, which in turn implies
that in any $\Sigma_{CV}$ language, either onsets and nuclei are both epenthesized ($/V/ \rightarrow ^*V$, and /CVC/ $\rightarrow$ CV.C$^*$,.) or nuclei and codas are both deleted ($/V/ \rightarrow [V]$ and /CVC/ $\rightarrow$ CV.[C]). Presumably this is empirically false, so we must allow the two kinds of PARSE and two kinds of FILL preferences to each have their own place in the dominance hierarchy.
<table>
<thead>
<tr>
<th>Uninflected</th>
<th>Nom future</th>
<th>Future</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kentapal</td>
<td>ketapal-in</td>
<td>kentapal-ur</td>
<td>dugong</td>
</tr>
<tr>
<td>kētār</td>
<td>kētār-in</td>
<td>kētār-ur</td>
<td>river</td>
</tr>
<tr>
<td>miyarq</td>
<td>miyarq-in</td>
<td>miyarq-ur</td>
<td>spear</td>
</tr>
<tr>
<td>yarpur</td>
<td>yarpur-in</td>
<td>yarpur-ur</td>
<td>snake, bird</td>
</tr>
<tr>
<td>yarman</td>
<td>yarman-in</td>
<td>yarman-ur</td>
<td>horse</td>
</tr>
<tr>
<td>pirņen</td>
<td>pirņen-in</td>
<td>pirņen-ur</td>
<td>woman</td>
</tr>
<tr>
<td>(B)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mela</td>
<td>mela-n</td>
<td>mela-ŋ</td>
<td>sea</td>
</tr>
<tr>
<td>wanka</td>
<td>wanka-n</td>
<td>wanka-ŋ</td>
<td>arm</td>
</tr>
<tr>
<td>kūŋa</td>
<td>kūŋa-n</td>
<td>kūŋa-ŋ</td>
<td>groin</td>
</tr>
<tr>
<td>gukat</td>
<td>gukā-n</td>
<td>gukā-ŋ</td>
<td>water</td>
</tr>
<tr>
<td>katu</td>
<td>katu-n</td>
<td>katu-ŋ</td>
<td>child</td>
</tr>
<tr>
<td>gawt</td>
<td>gawt-n</td>
<td>gawt-ŋ</td>
<td>wife</td>
</tr>
<tr>
<td>kenje</td>
<td>kenje-n</td>
<td>kenje-ŋ</td>
<td>wire</td>
</tr>
<tr>
<td>nëto</td>
<td>nëto-n</td>
<td>nëto-ŋ</td>
<td>skin</td>
</tr>
<tr>
<td>pape</td>
<td>papi-n</td>
<td>papi-ŋ</td>
<td>father's mother</td>
</tr>
<tr>
<td>tempe</td>
<td>tempe-n</td>
<td>tempe-ŋ</td>
<td>mother's father</td>
</tr>
<tr>
<td>wite</td>
<td>wite-n</td>
<td>wite-ŋ</td>
<td>interior</td>
</tr>
<tr>
<td>(C)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yalulu</td>
<td>yalulun</td>
<td>yalulu-ŋ</td>
<td>flame</td>
</tr>
<tr>
<td>mayara</td>
<td>mayaran</td>
<td>mayara-ŋ</td>
<td>rainbow</td>
</tr>
<tr>
<td>wiwalu</td>
<td>wiwalan</td>
<td>wiwala-ŋ</td>
<td>bush mango</td>
</tr>
<tr>
<td>karikar</td>
<td>karikarin</td>
<td>karikari-ŋ</td>
<td>butter-dish</td>
</tr>
<tr>
<td>yilyyil</td>
<td>yilyyilin</td>
<td>yilyyil-ŋ</td>
<td>oyster (sp.)</td>
</tr>
<tr>
<td>(D)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jurara</td>
<td>Jurara-pa</td>
<td>Jurara-pa-ŋ</td>
<td>shark</td>
</tr>
<tr>
<td>galuk</td>
<td>galuk-in</td>
<td>galuk-ŋ</td>
<td>story</td>
</tr>
<tr>
<td>putuṭa</td>
<td>putukan</td>
<td>putuka-ŋ</td>
<td>short</td>
</tr>
<tr>
<td>murūnman</td>
<td>murūnman-ŋ</td>
<td>murūnman-ŋ</td>
<td>cullah</td>
</tr>
<tr>
<td>gauŋpuŋuŋ</td>
<td>gauŋpuŋuŋ-ŋ</td>
<td>gauŋpuŋuŋ-ŋ</td>
<td>termite</td>
</tr>
<tr>
<td>tipitipi</td>
<td>tipitipi-ŋ</td>
<td>tipitipi-ŋ</td>
<td>rockfish (sp.)</td>
</tr>
<tr>
<td>ḫapu</td>
<td>ḫapu-n</td>
<td>ḫapu-ŋ</td>
<td>older brother</td>
</tr>
<tr>
<td>munkmununjuk</td>
<td>munkmununjuk-n</td>
<td>munkmununjuk-ŋ</td>
<td>wooden axe</td>
</tr>
<tr>
<td>tumpuŋtumpuŋ</td>
<td>tumpuŋtumpuŋ-ŋ</td>
<td>tumpuŋtumpuŋ-ŋ</td>
<td>dragonfly</td>
</tr>
<tr>
<td>(E)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yukarapa</td>
<td>yukaran</td>
<td>yukarapa-ŋ</td>
<td>husband</td>
</tr>
<tr>
<td>wulkan</td>
<td>wulkan-n</td>
<td>wulkan-ŋ</td>
<td>fruit (sp.)</td>
</tr>
<tr>
<td>wuṭal ti</td>
<td>wuṭal-in</td>
<td>wuṭal-ŋ</td>
<td>meat</td>
</tr>
<tr>
<td>kantukan</td>
<td>kantukan-n</td>
<td>kantukan-ŋ</td>
<td>red</td>
</tr>
<tr>
<td>karwakarwa</td>
<td>karwakaran</td>
<td>karwakan-ŋ</td>
<td>wattled (sp.)</td>
</tr>
</tbody>
</table>

**CVC**

<table>
<thead>
<tr>
<th>Uninflected</th>
<th>Nom future</th>
<th>Future</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>ṭera</td>
<td>ṭer-in</td>
<td>ṭer-ŋ</td>
<td>'thigh'</td>
</tr>
<tr>
<td>yura</td>
<td>yir-in</td>
<td>yur-ŋ</td>
<td>'body'</td>
</tr>
<tr>
<td>nita</td>
<td>nɪ-in</td>
<td>nɪ-ŋ</td>
<td>'neck'</td>
</tr>
<tr>
<td>ṭjuita</td>
<td>ṭjil-in</td>
<td>ṭjul-ŋ</td>
<td>'hair'</td>
</tr>
<tr>
<td>marpa</td>
<td>mar-in</td>
<td>mar-ŋ</td>
<td>'hand'</td>
</tr>
<tr>
<td>papa</td>
<td>pur-ŋ</td>
<td>pur-ŋ</td>
<td>'excrement'</td>
</tr>
<tr>
<td>wauta</td>
<td>wun-in</td>
<td>wun-ŋ</td>
<td>'rain'</td>
</tr>
<tr>
<td>kantua</td>
<td>kan-in</td>
<td>kan-ŋ</td>
<td>'grass'</td>
</tr>
<tr>
<td>njuha</td>
<td>njit-in</td>
<td>njith-ŋ</td>
<td>'fire'</td>
</tr>
</tbody>
</table>
Connectionism for Linguistics

Paul Smolensky

LSA Linguistic Institute
July 17, 1991

Collaborators:
Géraldine Legendre (Linguistics, Colorado)
Yoshiro Miyata (Cognitive Science, Chukyo)
Alan Prince (Linguistics, Brandeis)

Connectionist Computation
What is it? Low level characterization:
Data: activation values
Processing: spreading activation
Learning: associationist strength modification with experience

Why is it? Cognitive motivations:
Biological neural networks: abstract out a computational description
Psychological models of:
Perception
Memory
Associationist processing
Empiricist learning

Computation & linguistics
Psycholinguistics
Neurolinguistics
On-line processing
Acquisition
The form of grammar
If cognition is connectionist computation, what are the implications for grammar?

Relation between connectionist and symbolic computation
Eliminativism: Connectionist replaces symbolic
Implementationalism: Connectionist implements symbolic
Revisionism: Connectionist revises/enriches symbolic

[A major research program is required to pull this off; multiple levels.]

Grammar
Eliminativism: Grammar is a fiction.
Implementationalism: Grammar is unchanged.
Revisionism: Grammar inherits new kinds of rules & representations from connectionist computation

Requires high-level characterization of connectionist computation [which in turn requires a major research program to develop]
Computing done by linguist vs. computing done by speaker/hearer.

Examples of new ways of computing/thinking:
1. Continuous representations
2. Similarity-based processes
3. Parallel soft constraint satisfaction
4. Quantitative measures of well-formedness
5. Optimality as main organizing principle

Are these good ways of computing/thinking? You tell me.

Plan for lectures:
For \( p \in \{1,2,3,4,5\} \)
A. Develop high-level principle \( p \) of connectionist computation.
B. Propose a way of importing it into linguistics.
C. Develop an example application.

Logical progression of principles: \( 1 \rightarrow 5 \) (lowest, ‘most connectionist’ \( \rightarrow \) highest, ‘most symbolic’)

Conservative presentation of linguistic applications: \( 5 \rightarrow 1 \) — wins.

Today: 5; Next week: 4 - 1

5. Optimality as an organizing principle of grammar

A. Connectionist principle:

Harmony maximization. In a central and interesting class of connectionist networks, activation flow creates from an input a pattern of activity that [locally] maximizes a numerical well-formedness measure, the Harmony:

\[ \Sigma_i \Sigma_j a_i w_{ij} a_j \]

Linguistic interpretation:
connectionist input \( \rightarrow \) linguistic ‘input’ to grammar (string of words, phonemes)
connectionist well-formedness \( \rightarrow \) linguistic well-formedness
pattern of activity arising from an input \( \rightarrow \) structural description of the input
pattern of connections \( \rightarrow \) grammar

Here, all grammatical representations will be symbolic: traditional, unrevised. The rules too will be traditional in form, but their interpretation and interaction will be new. (Conservatism.)
B. Proposed linguistic principle:

Optimality. A grammar is a preference relation $>$ among structural descriptions; it assigns to any input that structure which is optimal (preferred by $>$ to all other structures).

The preference relation $>$ among structures is built up compositionally from preference relations among the substructures or subdimensions from which structures are built.

The basic preference relations and their means of combination are drawn from a universal repertoire.

Linguistic-internal motivation:
Grammars are more explanatory when they explicitly state well-formedness constraints rather than describing procedures to satisfy those constraints.
(Prefer principles to rewrite rules/transformations.)

Phonology has not made the progress towards this goal that syntax has — because the constraints are so often of the form, 'prefer … if possible' rather than 'obey … or be ill-formed'.
(But re:syntax, cf. Sadock's constraints on inter-component discrepancies.)

C. Applications in phonology [for many more, cf. Prince & Smolensky, LINS 247]

a. Berber syllabification

[insert brief statement of problem]

Idea: Dell—El-Medlaoui algorithm is trying to build the best syllables possible, as defined primarily by the goodness of their peaks. Replace the algorithm with a formal statement of this well-formedness metric.

More sonorous segments make better syllable peaks:
align sonority prominence hierarchy 8$\Rightarrow$7$\Rightarrow$...$\Rightarrow$1 with syllable structure prominence hierarchy: peak=non-peak

$(8,\text{peak}) > (7,\text{peak}) > ... > (1,\text{peak})$.

Given a syllabification $s$, let $M_s = \{M_x\}$ be the (multi-)set of peaks in $s$.

Let first($M_s$) = best peak of $M_s$, and let rest($M_s$) = $M_s \setminus \text{first}(M_s)$.

Then recursively define:

$M_x > M_y$ iff

first($M_x$) $>$ first($M_y$)

or

first($M_x$) = first($M_y$) and rest($M_x$) $>$ rest($M_y$)
Note: This means of composing the ordering $>$ on peaks to an ordering $>$ on syllabifications is a generalization of alphabetical ordering of strings: reinterpret first as ‘left-most element’. It turns out to be useful quite generally in the theory, so we dub it Generalized Lexicographic Ordering, GLO. To define a GLO $>$, we must specify the sets $M_s$, the function first, and an ordering $>$ of individual elements of $M_s$.

b. Lardil

A Strict Domination Grammar is a preference relation $>$ defined from a sequence of Boolean predicates over possible structural descriptions

$$p_1 \gg p_2 \gg \ldots \gg p_n$$

by the following case of GLO:

For a given structure $s$, let $M_s = \{ \pm p_1, \pm p_2, \ldots \}$ where the sign of $p_k$ indicates whether $s$ satisfies predicate $p_k$ (+) or not (−). Let first($M_s$) = first element of $M_s$. On each predicate separately, $\pm p_k$ (preference satisfaction) $>$ $\mp p_k$ (preference violation).

Note: This is a form of markedness theory. Each ‘$\mp p_k$’ is a kind of ‘mark’ on the structure, and the ordering $>$ gives a formal means of deciding which structure as a whole is least marked, by examining individual marks.

For clarity, we write ‘Prefer $p_k$’ for $p_k$ and ‘Avoid $q_k$’ when $p_k = \sim q_k$.

The Lardil preferences:

- $p_1$: Avoid subminimal words [less than one foot].
- $p_2$: Avoid non-final defective syllables.
- $p_3$: Avoid complex consonants in coda position [allowed: one non-back coronal].
- $p_4$: Prefer filled onsets.
- $p_5$: Prefer free final V.
- $p_6$: Avoid defective syllables.
- $p_7$: Avoid free material.
- $p_8$: Prefer coincidence of stem & syllable boundaries.