Constraint Interaction: A Lingua Franca for Theories of Language

Matt Goldrick
Department of Linguistics
Northwestern University

Phonology: Grammar vs. Process
- Generative grammars: Specify a function mapping underlying forms to surface forms.
  - Formalism: Optimality Theory; function is specified by the interaction of a set of ranked, violable constraints.
- Psycholinguistic theories: Specify how long term memory representations of phonological form are used to construct context-specific utterances plans (e.g., associating segments to prosodic structure).
  - Formalism: Connectionist models; mechanism is implemented via spreading activation between simple representational units.

Phonology as Constraint Interaction
- Grammars and psycholinguistic processes are often regarded as categorically distinct...
  - Historically focused on different types of data
  - Stated in radically different forms.
- These surface disparities mask a commitment to the underlying principle of constraint interaction.

Revealing Common Principles
Outline of Talk
- Mathematical analysis of Warker and Dell’s connectionist account of phonological processing
  - Characterize network processing in terms of constraint interaction.
  - How do these constraints allow the network to account for the data reviewed by Dell?
- Comparison with grammar-based approaches
  - Brief review: stochastic OT
  - Broadly similar forces at work in two approaches
- Theoretical debates are not grammar vs. process
  - What is the precise nature of constraints and their interactions?

Connectionist Phonology: Warker and Dell’s Syllabification Model

Where are the Constraints?
- “One thing that connectionist networks have in common with brains is that if you open them up and peer inside, all you can see is a big pile of goo. Internal organization is obscured by the sheer number of units and connections.”
  (Mozer & Smolensky, 1989:3)
- Reveal constraints by considering a simplified version of network and learning algorithm that share core principles of Warker and Dell (W&D)’s complex model.
Extracting Constraints from the Network

- W&D’s model is one specific instantiation of the covariant learning hypothesis (Van Orden, Pennington, & Stone, 1990).
  - Learning involves strengthening connections between covarying representational units or sets of representational units.
- Technique here: analyze a simpler instantiation of covariant learning (Hebbian learning) to reveal how principles of model produce behavior.
  - Two-layer Hebbian network: No hidden units; weights just reflect covariance of input/output units (not error correction).

Hebbian Network Behavior After Training

(1) $s_j^{(t)} = \alpha \left( \sum_p s_j^{(p)} F^{(p)}[p]|l| \right)$

Translation: the value of output unit $j$ for input pattern $t$ =
- The sum of outputs for all the training patterns $s_j^{(p)}$
- Contribution of each training pattern $p$ is weighted by its:
  - Frequency $F^{(p)}$ (weight: frequency)
  - Similarity to input pattern $O^{(p)}|l|$
- This sum is scaled by learning rate ($\alpha$) and squashed into the range (-1, 1) (using function $\sigma$)
* Crucial observation: Network output reflects the ensemble of training patterns, weighted by frequency and similarity.

(Plant et al., 1996; see Appendix 1)

3 Classes of Frequency + Similarity Effects

- Assume that training patterns all have outputs of +1 or −1.
- Rewrite (1) to show the network output for trained pattern $\alpha$.
- The degree to which network output reflects the correct output for $\alpha$ is related to:

  \[ s_j^{(\alpha)} = \alpha \left( \sum_{p:j \in \alpha} F^{(p)}[p]|l| \right) + \sum_{p:j \notin \alpha} F^{(p)}[p]|l| \]

  - Positively to the frequency of target pattern $\alpha$
  - Positively to the frequency of other patterns whose output $\alpha$ (weighted by their similarity to $\alpha$)
  - Negatively to the frequency of other patterns whose output $\neq \alpha$ (weighted by their similarity to $\alpha$)

*see also Plant et al., 1996, eq. 17

Constraint Interaction in Connectionist Networks

- Can conceive of these 3 effects as three constraint types.
- Network’s output is the determined by weighted interaction of these constraints.

  \[ s_j^{(\alpha)} = \alpha \left( \sum_{p:j \in \alpha} F^{(p)}[p]|l| \right) + \sum_{p:j \notin \alpha} F^{(p)}[p]|l| \]

  - Pressure to be correct (weight: frequency),
  - General support for target (weight: frequency, similarity)
  - Influence of “friends”;
  - General support for competitors (weight: frequency, similarity)
  - Influence of “enemies”;

*see also Stemberger (2004)

Connectionist Constraints + Empirical Data

- Syllable position effect (local constraint on errors)
  - “napkin” → “napkin” not “napkin”

  \[ s_j^{(\alpha)} = \alpha \left( \sum_{p:j \in \alpha} F^{(p)}[p]|l| \right) + \sum_{p:j \notin \alpha} F^{(p)}[p]|l| \]

  - First network constraint: be similar to trained pattern.
  - Note: in W&D’s simulations, all patterns have the same frequency (so effect is constant across targets).
  - (Some indication of frequency effects in their experimental data.)

Connectionist Constraints + Empirical Data

- Learning of first order constraints
  - E.g., /k/ is restricted to onset.
  - Speech errors: /k/ overwhelming in onset.

  \[ s_j^{(\alpha)} = \alpha \left( \sum_{p:j \in \alpha} F^{(p)}[p]|l| \right) + \sum_{p:j \notin \alpha} F^{(p)}[p]|l| \]

  - Training: exposure to syllables that are friends, no syllables that are enemies.
  - Increases strength of second constraint relative to third.
**Connectionist Constraints + Empirical Data**

- **Constraint complexity**
  - Second order constraints: if vowel is /æ/, /ɜ/ is in onset; if vowel is /il/, /ɛl/ is in coda.
  - Take longer to learn than first order constraints (Wark & Dell, in prep.).

**Connectionist Constraints + Empirical Data**

- **Second-order constraints are harder to learn because training strengthens general support for competitors.**
  - Since input representation has position-independent phonemes, enemies are presented during training.

**Connectionist Constraints + Empirical Data**

- **Connectionist Networks + Constraints**
  - Behavior of network can be described in terms of constraint interaction.
    - **Local position effects.**—Constraint pressuring output to reflect target.
    - **First order constraint learning.**—Strengthening of general constraint that prefers target representation.
    - **Slower second order constraint learning.**—Crosstalk due to position-independent phoneme representations interfere with learning.
    - **Relevant vs. irrelevant dimensions.**—Constraints are stated over input/output representations which have a certain structure.

**Connectionist Constraints + Empirical Data**

- **Relevant vs. irrelevant dimensions**
  - First order learning influenced by featural similarity (Goldrick, 2004)
  - Second order constraints using speaker voice (Onishi, Chambers & Fischer, 2002), speech rate (Dell) not learned.

**Connectionist Constraints + Empirical Data**

- **Overlap of patterns influences their effect.**
  - Features are part of representation; speaker voice, speech rate are not.

**Gratamatical Constraints + Speech Production**

- **Optimality Theory:** Constraint-based perspective on grammar.
  - Stochastic OT: allows constraint rankings to vary, allowing the grammar to model probabilistic behavior (e.g., speech errors).

- **Ex:** English variable R-deletion (Austin & Cho, 1998).
  - Proposed structure of grammar: In dialects with variable deletion, certain faithfulness and markedness constraints are unranked with respect to each other.
  - Stochastic component: A total ordering of constraints is randomly chosen at evaluation.

- **Ex.** Input: /Homer/. If Max, NoCoda are not ranked, yields variation.
  - Ranking 1: Max >> NoCoda. Output: [Ho.mer.]
  - Ranking 2: NoCoda >> Max. Output: [Ho.me.<r>]}
Grammatical Constraints + Speech Production

- More recent proposals have allowed OT grammars to specify more complex probability distributions (e.g., Boorsma & Hayes, 2001).
  - Stochastic OT grammars such as these have been used to model a variety of probabilistic behaviors—including experimentally elicited speech production data (Davidson, 2003).
  - See Appendix II for discussion of relationship of these models to other probabilistic models in computational linguistics.

Connecting Stochastic Grammars to Networks

- Consider target $\alpha$ and a single competitor structure $\beta$
  - Analogous to considering just output of a single unit [either +1 or −1] in connectionist network.
- Three types of constraints
  - Faithfulness ($\alpha$): Constraints preferring preservation of $\alpha$ (i.e., the input) in the output.
  - Markedness ($\alpha$): Constraints disfavoring structure $\alpha$ in the output.
  - Markedness ($\beta$): Constraints disfavoring structure $\beta$ in the output.

Relating Rankings to the Network

- Elements of the network equation can be associated with the probability of certain constraint rankings.
  - Note: probabilities of these rankings are not additive

$$s_\alpha^{(1)} = \mathcal{O} \left( s_\alpha^{(0)} + \sum_{\beta} F(\alpha) \prod_{\beta \neq \alpha} F(\beta) \right)$$

- $Pr(\text{Mark}(\alpha) >> \text{Mark}(\alpha))$: Preference for preserving target.
- $Pr(\text{Mark}(\beta) >> \text{Mark}(\alpha))$: General preference for target (= dispreference for competitors.)
- $Pr(\text{Mark}(\alpha) >> \text{Mark}(\beta))$: General preference for competitors.

Theoretical Debate: Nature of Constraints

- Both proposals rely on the interaction of three broad classes of constraints.
  - Distinctions therefore lie not in the framework, but in the detailed structure of constraints and their interactions; e.g.:
    - What representations are constraints stated over?
    - Is constraint domination strict or numerically weighted?
- Contrasts between (stochastic) generative grammars and psycholinguistic theories are not arguments between two incompatible architectures, but are instead similar to contrasts between theories within each discipline.
  - Constraints are not inherently connectionist or grammatical.

Ex: “Grain” of Faithfulness

<table>
<thead>
<tr>
<th>Theoretical Claims (two extremes)</th>
<th>Grammatical Realization</th>
<th>Connectionist Realization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coarse grain: Pressure to preserve target features is equal for all inputs</td>
<td>Ident (voice): For segments in any input, output segments in correspondence must have identical values for [voice]. (McCarthy &amp; Prince, 1999)</td>
<td>Activation of output units specifying voicing is constant across all inputs. (Wheeler &amp; Troske, 1997)</td>
</tr>
<tr>
<td>Fine grain: Pressure to preserve target features is input-specific</td>
<td>Exemplar models: Idiosyncratic properties of particular previous productions influence output. (Peter Hanssens, 2002)</td>
<td>$F^{(D)}$: Network weights are tied to frequency of individual training patterns.</td>
</tr>
</tbody>
</table>

Towards a Common Formal Vocabulary

- Constraint interaction provides a rich, common vocabulary for both linguists and psycholinguists to describe the structure, use, and acquisition of our knowledge of language.
- Recognizing common principles across these domains may help focus attention on the true substance of theoretical debates.
  - E.g., not grammars vs. networks, but the degree to which faithfulness is coarse vs. fine grained.
Appendix I: Derivation of Eq. (1)  
(based on Plaut et al. 1996)

- (A1) Hebbian learning rule: $\Delta w_{ji}^{[p]} = E s_j^{[p]} s_i^{[p]}$
- (A2) At end of training: $w_{ji} = E \sum_p s_j^{[p]} s_i^{[p]}$
- (A3) Network output: $s_j^{[1]} = O \left( \sum_i w_{ji} s_i^{[1]} \right)$
- (A4) Substitute A2 in A3: $s_j^{[1]} = O \left( \sum_{j'} \sum_i w_{ji} s_i^{[1]} s_j^{[1]} \right)$

Appendix II:  
Relationship to Computational Linguistics

- Stochastic OT models are closely connected to a family of models in computational linguistics, including:
  - Gibbs distributions
  - Maximum Entropy Models
  - Conditional Random Fields  
    (Johnson, 2002; see Goldwater & Johnson, 2003; Wilson, 2005, for applications to phonology).

- Note—not identical to stochastic OT: In these systems, constraints have real weights, so strict domination can only be approximated.
  - See Jäger & Rosenbach (in press) for further discussion.

Appendix II:  
Constraints in Gibbs Distributions

- General class of models based on Gibbs distributions: (Johnson, 2002): Probability candidate $j$ is output given input $t$: 
  $$\frac{1}{Z[t]} e^{\beta \sum \lambda_p f_p (J_t)}$$

Normalization term (= sum of constraint violations over all possible outputs). Converts weighted sum of constraint violations to probabilities.

All: Derivation of Eq. (1)

- (A5) Rearrange (A4) terms: $s_j^{[1]} = O \left( \sum_{j'} \sum_i w_{ji} s_i^{[1]} s_j^{[1]} \right)$
- (A6) Let $O^{[p][1]}$ stand for overlap of patterns ($\in$ dot product).
- (A7) Assume patterns can be presented multiple times during training. Let $F$ stand for the number of times each pattern is presented.
- Substituting (A7), (A6) into (A5) yields (1).

All: Constraints in Gibbs Distributions

- Assume constraints come in two types.
  - Faithfulness constraints penalize deviation from “target” outcome:
    - Faith(target candidate, [input]) = 0
    - Faith(competitor candidate, [input]) > 0
  - Markedness constraints penalize certain structures regardless of input.
All: Constraints in Gibbs Distributions

- Consider $\text{j}_x$ = target outcome and $\text{j}_y$ a single competitor (as in stochastic OT analysis above). Rewriting (A9) shows the log relative probability of producing target outcome is then a function of the same three constraint types as above:

$$
\begin{align*}
\text{Pressure to be correct (i.e., pressure to not produce incorrect outcome).} \\
\text{General pressure to be } \alpha \\
\text{General pressure to be } \beta
\end{align*}
$$

Note: Unlike stochastic OT, forces are additive just as in connectionist net.

References


