Poverty of the Stimulus? A Rational Approach

Perfors, Tenenbaum and Regier



Outline of PTR's Project

- Offer an abstract model of PoS using formal language
- Examine the relationship between learning and structure
- Discuss problem of simplicity versus fit for modeling
- Explain how hierarchical Bayesian model addresses structure.
- Show how hierarchical phase structure grammars gain traction over linear grammars.

The Primary Intuition of PoS

- Statistical models do not engage with the primary intuition...raised by PoS argument:
  - Language has a hierarchical structure.
  - It uses symbolic notions like syntactic categories and phrases that are hierarchically organized within sentences, which are recursively generated by a grammar.

Shortcomings of Non-hierarchical Models

- Connectionist and n-gram models are difficult to understand analytically (in linguistic or psychological terms).
- Prediction of next word in sequence does not examine grammar itself or what precisely is learned and why.

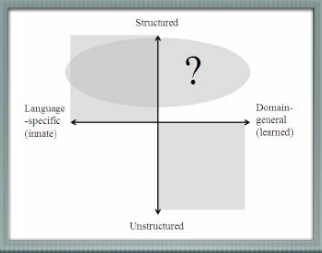
PTR's Goal: Address Structure through HBM

- Framework: Ideal Learnability Analysis - Theoretical not actual.
- Structure Matters: How to Capture? Use HBM!
  - "Our question is not whether a learner without innate language-specific biases *must* be able to infer that linguistic structure is hierarchical but rather whether it is *possible* to make that inference.
  - Results: Inference is possible. Let's see how.

PTR's Different Approach to Learning

- Two Fundamental Questions of Linguistic Knowledge
  - Do human learners have innate language-specific knowledge?
  - To what extent is linguistic knowledge based on structured representations such as generative phase-structure grammar?
- PTR argue that the two questions are not necessarily related although previous researchers have confounded the two issues.

Structure Need Not Be Innate



Theoretical Landscape for Language Acquisition
If language has structure, models must explain how it emerges.

PTR Offer a Formal Model of PoS

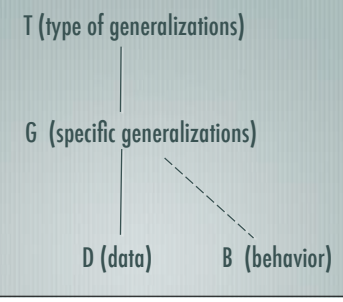
- (i) Children show a specific pattern of Behavior *B*
- (ii) A particular generalization *G* must be grasped in order to produce behavior *B*.
- (iii) It is impossible to reasonably induce *G* simply on the basis of the data *D* that children receive.
- (iv) Therefore, some abstract knowledge *T*, limiting which specific generalizations *G* are possible, is necessary.

Two Possible Models: Linear or Hierarchical

- 1a. Linear: Move first occurrence of auxiliary in the declarative to front to form interrogative.
- 1b. Hierarchical: Move auxiliary from the main clause of declarative to beginning to form interrogative.
  - The man is hungry. ⇨ Is the man hungry?
  - The man who is hungry is ordering dinner. ⇨ **?**
- Is the man who (is) hungry ordering dinner?

What Explains Behavior? A Model

Graphic Model of PoS Intuition



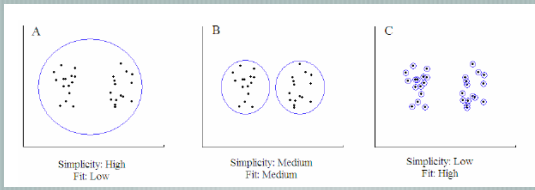
Comments on Logic of PoS

- It appears that some abstract knowledge *T* constrains G.
- *T* does not need to be specific to language. Other things (domain-general learning, memory etc.) could also constrain *T*.
- In this model, what matters is that higher order knowledge T is **necessary.**

Move to Innateness by PoS

- (i) Children show a specific pattern of Behavior *B*
- (ii) A particular generalization *G* must be grasped in order to produce behavior *B*.
- (iii) It is impossible to reasonably induce *G* simply on the basis of the data *D* that children receive.
- (iv) Therefore, some abstract knowledge *T*, limiting which specific generalizations G are possible, is necessary.)
- *T* could not itself be learned, or could not be learned before the specific generalization *G* is known.
- (vi) Therefore, *T must be innate.*

Simplicity versus Fit: Tradeoffs



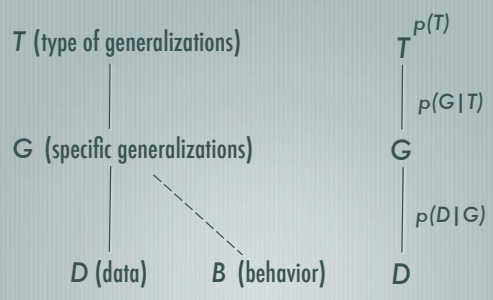
Model specified by location size of n Circles for:
(A) 1; (B) 2; (C) 30.

Ideally we would like to find Model that approximates B

PTR's finding (preview)

- "It may require less data to learn a higher-order principle *T* - such as the hierarchical nature of linguistic rules - than to learn every correct generalization *G* at a lower level, e.g. every specific rule of English."
- Thus we would want a type of Grammar that fits the data (corpus) well yet also offers a way to generalize to novel output without imposing unnecessary complexity (simplicity is preferred)

Transforming Intuition of PoS to HBM



- "Picking the grammar that best fits a corpus of child-directed speech [is] an instance of Bayesian model selection."
- "Model assumes that linguistic data is generated by picking a type of grammar *T*, then selecting as an instance of that type a specific grammar *G* from which the data *D* is generated."
- This gives us Bayes' rule:
  - $p(G, T \mid D) \propto p(D \mid G,T)p(G \mid T)p(T)$

Assumptions of HBM

- Learner can effectively search over joint space of G and T for grammars that maximize Bayesian scoring criterion.
  - What is the evaluation metric?
- Ideal learner will learn a given G, T pair rather than alternative G′, T′ if the former has a higher posterior probability than the latter.
- "We compare grammar types by comparing the probability of the best specific grammars *G of each type.*"

How Hierarchy Can Be Inferred

- Analytically come up with different ideal-types of grammar. All of these grammars successfully parse relevant corpora. These are hand-designed and are intended to offer the best case for each type. These form the generic possible types *T*, from which the learner can choose:
  - Flat (comprehensive list of memorized sentences)
  - Regular (Chomskian, akin to a Markov Chain)
  - Hierarchical Context-Free (Hierarchical Bayesian)

Probabilistic Grammars and Generalizability

- "All grammars are probabilistic, meaning that each production is associated with a probability and the probability of any given parse is the product of the probabilities of the productions involved in the derivation."
- Grammar efficacy depends on the available data. To test, corpus is stratified into 6 levels to see how well the models work given increasingly rich levels of input. Recall that all three grammar types can parse the full corpus. The question is how well does each generalize?

Stratifying the Corpus into 6 Levels

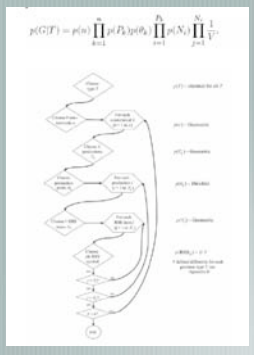
- Preference for a grammar is based on available data.
- PTR break corpus into 6 levels (analytic, not natural) based on frequency and complexity. Level 1 is lowest and has most frequent tokens, level 6 is highest (full corpus). These levels do not map to a child learner's ability.
- Recall that evaluation model's goodness is combination of simplicity and fit. After Chomsky, simplicity is preferred a priori. PTR wish to show how simplicity is preferred computationally.

Uses of Hierarchical Bayesian Models

- Use HBM to determine which grammars are chosen
  - >> $p(G|T) = p(P)p(n)\prod_{i=1}^{P} p(N_i)\prod_{j=1}^{N_i}\frac{1}{V}$
- Use HBM to measure how well grammar produces the test corpus.
  - >> $\log(p(D|G,T)) = \sum_{i=1}^{k}\log(p(S_i|G,T))$

- How Grammars are Generated (from Flat to PCFG)
  - Prior probability of a grammar reflects its complexity
  - Each grammar is selected from the space of all grammars of that type. More complex grammars are those that result from more (and more specific) choices.
  - Complex models require more free parameters to specify
  - PTR's model starts from scratch: choice is equally weighted between available types *T*, here designated as Flat, Regular or Context Free.

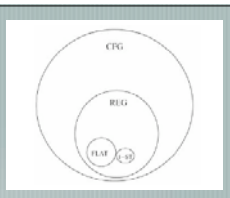
Flowchart of Grammar Generation using HBM



Choices (in order):
1. Grammar Type (flat, Reg, PCFG)
2. Number of non-terminals, productions, number of right-hand-side items each production contains.
3. For each item, a specific symbol is selected from the set of possible vocabulary (non-terminals and terminals).
4. This gives the prior probability for a grammar with *V* vocabulary items, *n* non-terminals, *P* productions and *N* symbols for production *i*.

- What? Explain the chart in English please!
- Subsets of grammars can be generated by the different grammar types and are not mutually exclusive. However, a particular grammar would have different prior probabilities under different types.
- A grammar with a certain number of productions, each of a certain size, has the **highest** probability if it can be generated as a **flat** (or one-state, not discussed here) grammar, next as a regular grammar and lastly as a CFG.

- Visually it looks like this:



- Flat grammars are subset of Regular grammars which are a subset of context-free grammars.
- Ceteris paribus, one has to make fewer choices in order to generate a specific regular grammar from the class of regular grammars than the class of context-free grammars.

Two-Component Model of Language

C1: Grammar assigns a probability distribution over potentially infinite set of syntactic forms that are accepted in language.

C2: Generates a finite observed corpus from the infinite set of forms produced by the grammar and can account for the characteristic power-law distributions found in language.

Model assumes separate generative processes for the allowable *types* of syntactic forms and for the frequency of specific sentence *tokens*.

Generating Types, Observing Tokens

Model maps well to psychological explanation: language users can generate syntactic forms by drawing on memory store or by consulting a deeper level of grammatical knowledge about how to generate all and only the legal syntactic forms.

"Grammar learning is based on how well each grammar accounts for the sentence types rather than on the frequencies of different sentence forms."

- Likelihood of a grammar is interpreted as fit to data and is dependent on quantity of data observed.
- HBM uses a probabilistic preference for the most specific grammar (size principle in concept learning), akin to a subset principle:
  - if a learner only sees positive examples of the target grammar, posits a single hypothesis at any one time and only learns from errors (current hypothesis fails to parse current sentence), and if learner posits a hypothesis which generates a superset of the true grammar the mistake will never be corrected and the true grammar will never be learned.
- Bayesian model approximates subset principle as data approach infinity as weight of likelihood increases while prior remains fixed.

Explaining Prior, Likelihood and Posterior Scores (Table 2) Part I

- Simplicity versus Compression explains transition from linear to hierarchical model preference.
- Lower levels of data input, linear models favored. At higher levels, CFG offers compression advantages. Example (# of productions to parse corpus)*:
  - Level 1: Flat= 8 ; Reg= 15; CFG= 20, whereas
  - Level 6: Flat= 2336; Reg= 169; CFG =69

* Actual results from the longer PTR paper but they are essentially the same.

Part II: Transitions

- Flat grammars have highest likelihood since they are a list of all the sentence types. But they do not generalize at all. Moreover, as input increases, they scale poorly and require too many productions.
- Regular grammars have high likelihood as well but generalize poorly.
- CFG have lower likelihoods but because they generalize well (predict unseen sentence types), they have less probability mass to predict observed sentences.
- Reviewing corpus again, as levels increase (4 to 6) data contain more recursive productions, which increase # of productions for linear grammars. CFG gains traction (greater posterior probability).

Percentage of Full Corpus Parsed Given Level of Data Input - Sentence Types

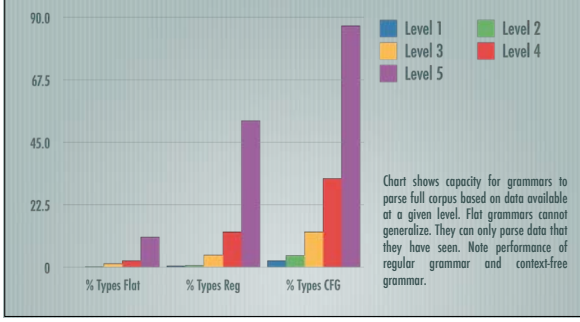
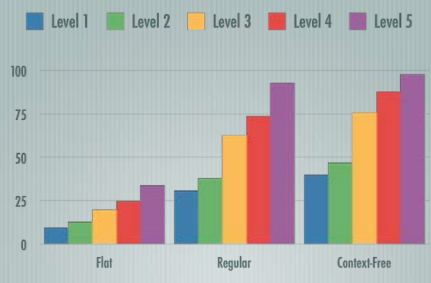


Chart shows capacity for grammars to parse full corpus based on data available at a given level. Flat grammars cannot generalize. They can only parse data that they have seen. Note performance of regular grammar and context-free grammar.

Percentage of Full Corpus Parsed Given Level of Data Input - Sentence Tokens



Part III: CFG>REG for AUX-Fronting

- CFG > REG > Flat
- CFG: succeeds because it has seen simple declaratives and interrogatives, allowing it to add productions in which the interrogative production is an aux-initial sentence that does not contain the auxiliary in the main clause.
- Regular: Fails because it has no way of encoding whether or not a verb phrase without a main clause auxiliary should follow that NP. If there is no input that shows such a verb phrase did occur Regular grammar cannot produce it. To produce it would require specific examples, the kind of which Chomsky argues are absent or impossibly rare.

Conclusion

- PTR's corpus is best explained "using the hierarchical phrase structures of context-free grammars rather than the linear structures of simpler Markovian grammars."
- HBM confirms Chomsky's intuition of the structural nature of language. However, this structure need not be innate.
- HBM shows how the intuitive preference for simplicity can be incorporated into an effective probabilistic model.

- Examine question of whether 1) language capacities are innate and 2) the extent to which a particular capacity is domain-specific or domain-general.
- Reservations and assumptions:
  - Powerful learning mechanism needed (search over sum of all possible grammars.

- Powerful domain-general learning mechanism with few, weak innate biases or weak learning mechanism with stronger biases? PTR are agnostic but state that there must be an *a priori* bias to prefer hierarchical phase structure.
- What are the representational capacities of the child?
  - Children must have ability to represent both linear and hierarchical patterns (not necessarily confined to language alone).
  - Learners can represent different grammars types **as** types (distinguish kind). Here: flat, regular, context-free.
  - Learners confront a language which has explicit symbolic structure.
  - PTR use syntactic categories (sentence types) as input rather than lexical items. Is this too big a leap (assuming learner can form categories over which data are processed).

Final Thoughts

- In PTR's model the learner does not need to know a priori that language actually has hierarchical phase structure. The learner must only be aware that the possibility exists (thus the three types of *T*). The HBM model shows how a learner will move toward a hierarchical model as data input increases.