Introduction to Cognitive Science: Notes

X: Child and Computer Language Development

- Readings for this section: *Gopnik and Schulz 2004; *Zettlemoyer and Collins 2005.

Child and Computer Language Development

- The child's problem is similar to the problem of inducing a treebank grammar, but a little harder.
  - They have unordered logical forms, not language-specific ordered derivation trees.
  - So they have to work out which word(s) go with which element(s) of logical form, as well as the directionality of the syntactic categories (which are otherwise universally determined by the semantic types of the latter).

Child and Computer Language Development

- Children do not seem to have to deal with a greater amount of error than the Penn WSJ treebank has (McWhinnie 2005).
  - But they may need to deal with situations which support a number of logical forms.
  - And they need to be able to recover from temporary wrong lexical assignments.
  - And they need to be able to handle lexical ambiguity.

Computational Accounts

- Siskind (1995, 1996), Villavicencio (2002), and Zettlemoyer and Collins (2005) offer computational models of this process.
- Both theories make strong assumptions about the association of words with elements of logical form.
- Both make strong assumptions about universally available parametrically specified rule- or category- types, the latter in the form of a type hierarchy
- Both deal with noise and homonymy probabilistically.

Computational Accounts: Zettlemoyer and Collins

- Zettlemoyer and Collins' algorithm (UAI 2005) allows any contiguous substring of the sentence to be a lexical item. For a given logical form, the learner has to search the cross-product of the substring powerset of the string with the set of pairs of legal categories with elements of the substructure powerset of the logical form for categories that yield combinatory derivations that yield the correct logical form.
- Learning is via a log-linear model using lexical entries (only) as features and gradient descent on their weights, iterating over successive sentences of a corpus of sentence-logical form pairs.
- We can improve on this by
  - Directly generating the parses that UG supports for the sentence-meaning pair.
  - Building a full parsing model (necessary if we are to scale).

Zettlemoyer and Collins (Contd.)

- The algorithm as presented in 2005 learns only a very small rather unambiguous fragment of English, hand-labeled with uniquely identified database queries as logical forms, and an English specific inventory of possible syntactic category types in lieu of Universal Grammar.
- However, Siskind's and Villavicencio's results already tell us that the algorithm should work with multiple candidate logical forms.
- Similarly, their results show that a universal set of category types can be used without overwhelming the learner.

Zettlemoyer and Collins (Contd.)

- All of these models depend on availability to the learner of short sentences paired with logical forms, since complexity is determined by a cross-product of powersets both of which are exponential in sentence length.
- A number of techniques are available to make search efficient including use of a head-dependency parsing model.

The Generative Model

- We will assume that $P(D, I, S)$ is a generative model for an (exhaustive) parser, rather than the discriminative model of Zettlemoyer *et al.*.
- One advantage of generative models besides their closeness to competence grammar is that we can invert the parsing model to define the probability of an utterance given a meaning.

However, another difference between the child and standard treebank grammar-induction programs is that the child learns grammar *incrementally*, utterance-by-utterance.

Recomputing the model over the entire corpus so far, as each new sentence is encountered, is not only psychologically absurd, but computationally exponential.

Example

- The child thinks: $more' dog'$
- The Adult says: “More doggies!”
- Given the string “more dogs” paired with the logical form $more' dogs'$, and a mapping from semantic types onto syntactic type like S, NP, $S\backslash NP$ etc., the child can use the universal **BT**-based combinatory rules of CCG to generate
  - all possible syntactic derivations, pairing
  - all possible decompositions of the logical form with
  - all possible word candidates
- Learning a language is just learning its lexicon and a parsing model.

The Derivations

- CCG permits just three derivations for the new utterance “More doggies” , as follows:

(1) a.
$$\dfrac{\dfrac{\textsc{More}}{NP/N : more'_{((e,t),e)}} \quad \dfrac{\textsc{doggies}}{N : dogs'_{(e,t)}}}{NP : more'\,dogs'_{e}}\,{>}\;!$$

b.
$$\dfrac{\dfrac{\textsc{More}}{N : dogs'_{(e,t)}} \quad \dfrac{\textsc{doggies}}{NP\backslash N : more'_{((e,t),e)}}}{NP : more'\,dogs'_{e}}\,{<}\;!$$

c.
$$\dfrac{\textsc{More doggies}}{NP : more'\,dogs'_{e}}\;!$$

The Child's First Lexicon

- (2) The child's lexical candidates:

  more:= $\mathbf{NP/N} : \mathbf{more}'_{((\mathbf{e},\mathbf{t}),\mathbf{e})}$

  $N : dogs'_{(e,t)}$

  doggies:= $NP\backslash N : more'_{((e,t),e)}$

  $\mathbf{N} : \mathbf{dogs}'_{(\mathbf{e},\mathbf{t})}$

  more doggies:= $NP : (more'\, dogs')_e$

- A statistical model for these hypotheses can be learned using an incremental variant of the semi-supervised inside-outside (EM) algorithm (Pereira and Schabes 1992; Neal and Hinton 1999). We begin with a simplified model, representing probabilities as expected frequencies, then define the model we actually use.

Learning the Model for English

- In order to obtain an incremental algorithm, we represent the model as a vector of expected frequencies for each production p, defined as

(3) $fexp(p) = \sum_{s \in S} \sum_{i \in I} P(i|s) \sum_{d \in D} P(d|s,i).count(p,d)$,
where $P(d|s,i) = \frac{P(d)}{\sum_{d \in D} P(d)}$

The primary requirement for such a model is that learned information about seen events in a derivation should influence the probabilities assigned to unseen events.

- Thus, if the language only consists of sentences of the form “More X”, and the hundredth sentence is “More erasers”, where “erasers” is a previously unseen word, this sentence should not only make the learner a little more certain that “more” is a determiner meaning *more′*.
- It should also make them pretty sure that “erasers” is a noun, and *not* a determiner meaning *more′*.

Two Estimators for Expected Frequency

- We define two estimators for *fexp*.
- $Fexp_E$ is the expected frequency based on the present sentence and the possibilities of universal grammar alone. For simplicity we will assume the latter to be uniformly distributed, so that (3) reduces to the following, where $|D|$ is the number of derivations:

  (4) $fexp_E(p) = \frac{\sum_{d \in D} count(p,d)}{|D|}$

- $Fexp_M$ for a given interpretation i for sentence s is defined as follows, where P is the model estimated so far.

  (5) $fexp_M(p) = \sum_{i \in I} P(i|s) \sum_{d \in D} P(d|s,i).count(p,d)$

The Algorithm

- The model can be learned using the following incremental variant of the semi-supervised inside-outside (EM) algorithm (Pereira and Schabes 1992; Neal and Hinton 1999).
- Every new sentence s_n provides a set D_n of derivations parallel to (1), which defines the following:
  a. A (possibly empty) set of previously unseen productions involved in some derivation in D_i, including those involving novel lexical entries, that must be added to the model with cumulative *fexp* temporarily initialized to zero.
  b. (E-step): The set of all productions including those in a, whose cumulative *fexp* must be multiplied by $n-1$, incremented by $fexp_E$, and divided by n.
  c. (M-step):A further increment of $\frac{fexp_M - fexp_E}{n}$ (which may be negative) to the cumulative *fexp* for all productions involved in some derivation in D_i. I.e., replace the earlier estimate based on $fexp_E$.

The Algorithm

- Step b defines new values for the conditional probabilities for the rules in question, defining an intermediate model for calculating the a posteriori probabilities in step c.
- The further update c to the model defines the expected frequencies for the next cycle. The lexical probabilities for the relevant words in the lexicon given the new sentence can then be calculated using the model and definition (3), where $P(d|I,S)$ is the product of the probabilities of the productions it involves.
- (6) $P(d|I,S) = \prod_{p \in d} P(p|parent) \prod_{LEX(p) \in d} P(\phi, \sigma|\mu)$

This is just a probabilistic context-free grammar parser (PCFG). We actually use a head-dependency model (Collins 2003)

Normalizing Probabilities of Derivations

- The possibility of lexicalizing more than one element of the logical form in a single word means that the alternative derivations for a single logical form such as those in (1) for our running example and the first sentence “More doggies” may be of different lengths.
- Since generative models of the kind outlined above, based on the products of probabilities of rules, assign undue weight to short derivations, we must normalize the probabilities of lexical productions over the complexity of their logical forms.

Thus, the probability $P(\phi.\sigma|\mu)$ of the lexical productions in (6) is

(7) $P(\phi.\sigma|\mu) = \prod_{m \subset \mu} P(\phi,\sigma|m)$

- For example, the probability of derivation (1c) is not a third, but is the conditional probability of “more dogs” given *more*′*dogs*′ times that of “more dogs” given *more*′, times that of “more dogs” given *dogs*′—that is, $\frac{1}{3} \times \frac{1}{3} \times \frac{1}{3}$.

Probabilities of the Derivations

- Thus on the basis of the intermediate value $\frac{(0)fexp(0)+fexp_E}{1}$, the relative conditional probabilities $P(D|I,S)$ of the three derivations (1) are as follows:

(8) a $P(A|I,S) = P(r0|START) \times P(r1|NP : fa)) \times P_{lex}(more, NP/N|more') \times P_{lex}(doggies, N|dogs') = \frac{1 \times 0.\dot{3} \times 0.\dot{3} \times 0.\dot{3}}{\sum_d P(d|I,S)}$

b $P(B|I,S) = P(r0|START) \times P(r2|NP : fa)) \times P_{lex}(doggies, NP\backslash N|more') \times P_{lex}(more, N|dogs') = \frac{1 \times 0.\dot{3} \times 0.\dot{3} \times 0.\dot{3}}{\sum_d P(d|I,S)}$

c $P(C|I,S) = P(r0|START) \times P_{lex}(more\ \ doggies, NP|more') \times P_{lex}(more\ doggies, NP|dogs') = \frac{1 \times 0.\dot{3} \times 0.3 \times 0.\dot{3}}{\sum_d P(d|I,S)}$

$P(A|I,S) = P(B|I,S) = P(C|I,S) = 0.\dot{3}$

Child's First Parsing Model (Simplified)

- This means that the initial model can be calculated as follows:

(9)

Rule	$fexp(n-1)$	$\frac{(n-1)fexp(n-1)+fexp_E}{n}$	$fexp(n)$
r0. $START \rightarrow NP : fa$	0	1.0	1.0
r1. $NP : fa \rightarrow NP/N : f \quad N : a$	0	$0.\dot{3}$	$0.\dot{3}$
r2. $NP : fa \rightarrow N : a \quad NP\backslash N : f$	0	$0.\dot{3}$	$0.\dot{3}$
l1. $NP/N : more' \rightarrow$ more	0	$0.\dot{3}$	$0.\dot{3}$
l2. $NP\backslash N : more' \rightarrow$ doggies	0	$0.\dot{3}$	$0.\dot{3}$
l3. $N : dogs' \rightarrow$ doggies	0	$0.\dot{3}$	$0.\dot{3}$
l4. $N : dogs' \rightarrow$ more	0	$0.\dot{3}$	$0.\dot{3}$
l5. $NP : more' dogs' \rightarrow$ more doggies	0	$0.\dot{3}$	$0.\dot{3}$

The Child's First Lexicon

- Thus, we have the following updated probabilistic lexicon:

(10)

ϕ	σ, μ	f_{exp}	$P_{lex}(\sigma, \mu \mid \phi)$	$P_{lex}(\phi \mid \mu)$
more:=	$\mathbf{NP/N} : \mathbf{more}'_{((\mathbf{e},\mathbf{t}),\mathbf{e})}$	$0.\dot{3}$	$0.\dot{3}$	$0.\dot{3}$
	$N : dogs'_{(e,t)}$	$0.\dot{3}$	$0.\dot{3}$	$0.\dot{3}$
doggies:=	$NP\backslash N : more'_{((e,t),e)}$	$0.\dot{3}$	$0.\dot{3}$	$0.\dot{3}$
	$\mathbf{N} : \mathbf{dogs}'_{(\mathbf{e},\mathbf{t})}$	$0.\dot{3}$	$0.\dot{3}$	$0.\dot{3}$
more doggies:=	$NP : (more' dogs')_e$	$0.\dot{3}$	$0.\dot{3}$	$0.\dot{3}$

Early Overgeneration

- Since the word counts and conditional probabilities for “more” and “doggies” with them meaning $more'_{((e,t),e)}$ are all equal at this stage, the child may well make errors of overgeneration, using some approximation to “doggies” to mean “more”.
- However, even on the basis of this very underspecified lexicon, the child will not overgenerate “*doggies more”.
- Moreover, further observations, with further updates to frequency counts, will rapidly lower the estimated conditional probability of the spurious hypotheses concerning categories and substrings in comparison to the correct ones, indicated in bold type, as follows:

The Child's Second Sentence

- Let us suppose that the second utterance the child hears is "More cookies". There are again three derivations parallel to (1). The child can derive a new parsing model by adding new rules, updating expected frequencies for all rules in the new set of derivations, and recalculating a posteriori expected frequencies as described:

Prior Probabilities for the Three Possible Derivations

- On the basis of the intermediate value $\frac{(1)fexp(1)+fexp_E}{2}$,the length-weighted relative conditional probabilities $P(d|I,S)$ of the three derivations for "More cookies" parallel to (1) are as follows:

(11) a $P(A|I,S) = P(r0|START) \times P(r1|NP : fa)) \times P_{lex}(more, NP/N|more') \times P_{lex}(cookies, N|cookies') = \frac{1.0 \times 0.\dot{3} \times 0.\dot{3} \times 0.1\dot{6}}{\sum_d P(d|I,S)} = 0.42$

b $P(B|I,S) = P(r0|START) \times P(r2|NP : fa)) \times P_{lex}(cookies, NP\backslash N|more') \times P_{lex}(more, N|cookies') = \frac{1 \times 0.\dot{3} \times 0.1\dot{6} \times 0.1\dot{6}}{\sum_d P(d|I,S)} = 0.23$

c $P(C|I,S) = P(r0|START) \times P_{lex}(more\ cookies, NP|more') \times P_{lex}(more\ cookies, NP|cookies') = \frac{1 \times 0.\dot{3} \times 0.01\dot{6} \times 0.25}{\sum_d P(d|I,S)} = .35$

$P(A|I,S) \neq P(B|I,S) \neq P(C|I,S) \neq 0.\dot{3}$

The Child's 2nd Parsing Model (Simplified)

- (12)

Rule	$fexp(n-1)$	$\frac{(n-1)fexp(n-1)+fexp_E}{n}$	$fexp(n)$
r0. $START \rightarrow NP : fa$	1.0	1.0	1.0
r1. $NP : fa \rightarrow NP/N : f \quad N : a$	$0.\dot{3}$	$0.\dot{3}$	0.34
r2. $NP : fa \rightarrow N : a \quad NP\backslash N : f$	$0.\dot{3}$	$0.\dot{3}$	0.25
l1. $NP/N : more' \rightarrow$ more	$0.\dot{3}$	$0.\dot{3}$	0.34
l2. $NP\backslash N : more' \rightarrow$ doggies	$0.\dot{3}$	$0.1\dot{6}$	$0.1\dot{6}$
l3. $N : dogs' \rightarrow$ doggies	$0.\dot{3}$	$0.\dot{1}6$	$0.1\dot{6}$
l4. $N : dogs' \rightarrow$ more	$0.\dot{3}$	$0.1\dot{6}$	$0.1\dot{6}$
l5. $NP : more' dogs') \rightarrow$ more doggies	$0.\dot{1}$	$0.1\dot{6}$	$0.1\dot{6}$
l6. $NP : more' cookies' \rightarrow$ more cookies	0	$0.1\dot{6}$	0.17
l7. $NP\backslash N : more' \rightarrow$ cookies	0	$0.1\dot{6}$	0.11
r8. $N($cookies$) : cookies' \rightarrow$ cookies	0	$0.1\dot{6}$	0.24
l9. $N($more$) : cookies' \rightarrow$ more	0	$0.1\dot{6}$	0.11

The Child's Second Lexicon

- Thus, we have the following updated probabilistic lexicon:

(13)

ϕ	σ, μ	$fexp_{lex}(n)$	$P(\sigma,\mu\|\phi)$	$P(\phi\|\sigma,\mu)$
more:=	$\mathbf{NP/N : more'_{((e,t),e)}}$	**0.34**	**0.57895**	**0.57895**
	$N : dogs'_{(e,t)}$	$0.1\dot{6}$	0.26318	0.5
	$N : cookies'_{(e,t)}$	0.11	0.15789	$0.\dot{3}$
doggies:=	$NP\backslash N : more'_{((e,t),e)}$	$0.1\dot{6}$	0.5	0.385
	$\mathbf{N : dogs'_{(e,t)}}$	$0.1\dot{6}$	0.5	0.50
cookies:=	$NP\backslash N : more'_{((e,t),e)}$	0.11	$0.\dot{3}$	0.15789
	$\mathbf{N : cookies'_{(e,t)}}$	**0.24**	$\mathbf{0.\dot{6}}$	$\mathbf{0.\dot{6}}$
more doggies:=	$NP : (more' dogs')_e$	$0.1\dot{6}$	$0.\dot{3}$	$0.\dot{3}$
more cookies:=	$NP : (more' cookies')_e$	0.17	$0.\dot{3}$	$0.\dot{3}$

The Child's Second Lexicon

Notice that the expected frequencies in this table are not quite the same as those that would be obtained by recomputing f_{exp} over the entire corpus, as in standard batch EM.

- Nevertheless, at this point, the child is exponentially less likely to generate "doggie" when she means "more".
- Experimental sampling by elicitation of child utterances during such exponential extinction may well give the appearance of all-or-none setting of parameters like NEG-placement and *pro*-drop claimed by Thornton and Tesan (2006).
- This effect is related to the "winner-take-all" effect observed in Steels' 2004 game-based account of the very similar process of establishing a shared vocabulary among agents who have no preexisting language.

An Aside: A Statistically Sound Model

- We actually need a generative model that explicitly states the probabilities of the productions that are used in producing $\langle S, I, D \rangle$.
  - We model the probability of the syntactic derivation $P(D|START)$ using the PCFG type productions described before.
  - Each derivation gives a set of syntactic components $\underline{\sigma}$

An Aside: A Statistically Sound Model

- We can now approximate the conditional probability of the associated semantics as:
  - $P(\lambda_{lex}|\sigma_i, \Lambda) \approx \frac{1}{Z} P(\lambda_{lex}|\Lambda) * t(\tau_\sigma, \tau_\lambda)$
  - t is a binary function that checks that the types of the syntax and semantics are compatible.
  - Λ is a model of the semantics available to the system. We break the lexical probability up as follows:
  - $P(\lambda_{lex}|\Lambda) = \prod_{\lambda_c \in \lambda_{lex}} P(\lambda_{lex}|\lambda_c) \times P(\lambda_c|\Lambda)$
  - The $P(\lambda_{lex}|\lambda_c)$ terms allows us to penalise complex semantics that appear in the lexicon.
  - The $P(\lambda_c|\Lambda)$ terms allow us to penalise rare semantics.

An Aside: A Statistically Sound Model

- The probability of $\langle S, I, D \rangle$ is calculated as:

$$P(\langle S, I, D \rangle | START, \Lambda) = P(D|START) \times \prod_i P(\phi_i|\sigma_i, \lambda_i) P(\lambda_i|\sigma_i, \Lambda)$$

- The grammar must model the production probabilities $P(p|parent)$
- The lexicon must model $P(\lambda_{lex}|\lambda_c)$, $P(\lambda_c|\Lambda)$, $P(\phi|\sigma, \lambda)$
- Incremental updates are made to these probability distributions by calculating likelihoods given each new sentence (as before) and using Bayes's rule to update the posterior belief, which is then stored.
- In order to make this simple, the grammar rules are modelled using a Dirichlet prior and the lexical probabilities are modelled using Dirichlet Processes.
  - In both cases the likelihood is conjugate to the prior, so the updates are easy to perform

Later Development

- This effect is also all that is needed to explain the phenomenon of “syntactic bootstrapping” (Gleitman (1990)), where at a later stage of development, the child can learn lexical entries for words for which the corresponding concept is not salient, or is even entirely lacking to the child.
- In this connection it is important that the expected frequency of the non-English rule r2 is already dropping in comparison to r1.

The Real Point of Using CCG to Model Acquisition

- If children's exposure to language were merely confined to recitations of propositions they already had in mind, it would be a dull affair.
- It is not even clear why they would bother to learn language at all, as Clark (2004) points out in defence of a PAC learning model.
- We know from Fernald *et al.* (1989) and Fernald (1993) that infants are sensitive to interpersonal meanings of intonation from a very early age.
- In English, intonation contour is used to convey a complex system of information-structural elements, including topic/comment markers and given/newness markers (Bolinger 1965; Halliday 1967; Ladd 1996), and is exuberantly used in speech by and to infants.

For the child, it is this part of the meaning that is the whole point of the exercise.

Towards a More Realistic Syntax and Semantics

- For example, it is likely that the child's representation of the utterance "MORE DOGGIES! is more like (14), in which [S] represents speaker syupposition (contributed by the LL% boundary tone), ρ indicates a rheme or comment (contributed by the H* pitch-accents), * marks emphasis or kontrast (also contributed by the pitch-accents), and the category NP is "type-raised", indicated by the annotation $NP^{\uparrow}$:

(14)

$$\dfrac{\dfrac{\begin{array}{c}\text{MORE DOGGIES}\\ \text{H}* \quad \text{H}*\end{array}}{NP^{\uparrow}_{+,\rho} : \lambda p.p(*more'*dogs')} \quad \dfrac{\begin{array}{c}!\\ \text{LL\%}\end{array}}{X_{\phi}\backslash_{\star}X_{\pi,\eta} : \lambda g.\pi[\text{S}]\eta\, g}}{NP^{\uparrow}_{\phi} : [S]\rho\lambda p.p(*more'*dogs')}<$$

"Mummy makes the property afforded by more dogs common ground."

- Consider the child in a similar situation faced with the following utterance, from Fisher and Tokura (1996) as discussed in Steedman 1996:

(15)

You	LIKE H*		the doggies!	
		L		LL%
$S/(S\backslash NP)$	$(S\backslash NP)/NP$	$X_\phi \backslash_\star X_{\pi,\eta}$	$S_\phi \backslash (S_\phi/NP_\phi)$	
$: \lambda p.p\ you'$	$*like'$	$: \lambda g.\pi[S]\eta\ g$	$: [S]\eta\lambda q.q\ dogs'$	
	$S/NP : \lambda x.*like\ x\ you'$ (>B)			
	$S_\phi/NP_\phi : [S]\rho\lambda x.*like' x\ you'$ (<)			
	$S_\phi : ([S]\theta\lambda p.p\ dogs')([S]\rho\lambda x.*like' x\ you')$ (<)			
	$S : like'\ dogs'\ you'$			

"Mummy supposes what property the dogs afford to be common ground, Mummy makes it common ground it's me liking them."

- Fisher points out that the L boundary after the verb makes the intonation structure inconsistent with standard assumptions about surface constituency.

What CCG is Good For

- However, this intonation structure is isomorphic to the CCG derivation above, which delivers the corresponding theme/rheme information partition directly.
- Thus, here too, the availability of the full semantic interpretation, including information-structural information, directly reveals the target grammar.
- In this case, since the derivation requires the use of the forward composition rule, indexed $>\mathbf{B}$, the child gets information not only about the probability of the verb, the nominative, and the accusative categories of English, but also about the probability of applying the composition rule to the first two categories, the probability that the subject of "like" will be headed by "you", and its object be headed by "doggies".
- Thus, the child can build the entire parsing model in parallel with learning the grammar, including the long range dependencies.

Discussion

- Syntax is learned on the basis of preexisting semantic interpretations afforded by the situation of adult utterance, using a statistical model over a universal set of grammatical possibilities.
- The existence of the model itself helps the child to rapidly acquire a correct grammar even in the face of competing ambiguous semantics and error, without requiring the (empirically questionable) subset principle.
- The fact that the onset of syntactically productive language at the end of the Piagetian sensory-motor develomental phase is accompanied by an explosion of advances in qualitatively different “operational” cognitive abilities suggests that the availability of the statistical model has a feedback effect, allowing “Syntactic bootstrapping” of concepts to which the child would not otherwise gain access.

Parameters and Triggers Unnecessary

- The theory presented here somewhat resembles the proposal of Fodor 1998 as developed in Sakas and Fodor (2001) and Niyogi (2006) in treating the acquisition of grammar as in some sense parsing with a universal "supergrammar". As in that proposal, both parameters and triggers are simply properties of the language-specific grammar itself—in their case, rules over independently learned parts of speech, in present terms, lexical categories.
- Rather than learning rules in an all or none fashion on the basis of unambiguous sentences that admit of only one analysis, the present theory adjusts probabilities in a model of all elements of the grammar for which there is positive evidence for *all* processable utterances.

Against “Parameter Setting”

- In this respect, it resembles the proposal of Yang (2002). However it differs in eliminating explicit parameters.
- If the parameters are implicit in the rules or categories themselves, and you can learn the rules or categories directly, why should the child (or a truly Minimal theory) bother with parameters at all?
- For the child, all-or-none parameter-setting is counterproductive, as it will make it hard to learn the many languages which have inconsistent settings of parameters across lexical types and exceptional lexical items, as in German and Dutch head finality.
- Or consider English expressions like the following:

(16) Doggies galore!

“Galore” is the only phrase-final determiner in E. (stolen from Irish).

Conclusion

- Everything that we think of as cognitive, or having the property of *intentionality*, from stereo vision to semantics of discourse, is shaped by the primordial need to act in the world.
- The operations of composition **B** and type-raising **T** that form the basic of affordance and seriation in planning provide the basis for universal grammar beyond the lexicon.
- Thinking, Semantic Interpretation,and Understanding all take place in a dynamic context of goals and plans that is essentially pre-linguistic.

So Why Don't Apes have Productive Syntax?

- If composition and type raising are prelinguistic planning primitives that we share with other animals, what more is needed to support the language faculty?
- One candidate is modal and propositional attitude concepts—that is, functions over propositional entities. (We have so far glossed over the important fact that plans compose actions of type *state* → *state*, whereas syntax composes functions of type *proposition* → *proposition*.) These induce true recursion in conceptual structures and grammar via the grounded lexicon.
- There is no evidence that apes can attain the kind of theory of other minds that is required to support such concepts. Perhaps this is *all* they lack (Premack and Premack 1983;Tomasello 1999; Steedman 2002a,b; Hauser *et al.* 2002).
- If so, we need to know much more about the development of propositional attitude concepts, and their relation to planning and tool use around Piagetian sensory-motor developmental stage 6.

References

Bolinger, Dwight, 1965. *Forms of English*. Cambridge, MA: Harvard University Press.

Clark, Alexander, 2004. "Grammatical Inference and First Language Acquisition." In *CoLing Workshop on Psycho-computational Models of Human Language Acquisition*.

Collins, Michael, 2003. "Head-Driven Statistical Models for Natural Language Parsing." *Computational Linguistics* 29:589–637.

Fernald, Anne, 1993. "Approval and Disapproval: Infant Responsiveness to Vocal Affect in Familiar and Unfamiliar Languages." *Child Development* 64:657–667.

Fernald, Anne, Taeschner, T., Dunn, J., Papousek, M., Boysson-Bardies, B., and Fukui, I., 1989. "A Cross-language Study of Prosodic Modifications in Mothers' and Fathers' Speech to Infants." *Journal of Child Language* 16:477–501.

Fisher, Cynthia and Tokura, Hisayo, 1996. "Prosody in Speech to Infants: Direct and Indirect Acoustic Cues to Syntactic Structure." In James Morgan and

Katherine Demuth (eds.), *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*, Erlbaum. 343–363.

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

Gleitman, Lila, 1990. "The Structural Sources of Verb Meanings." *Language Acquisition* 1:1–55.

Gopnik, Alison and Schulz, Laura, 2004. "Mechanisms of Theory Formation in Young Children." *Trends in Cognitive Science* 8:371–377.

Halliday, Michael, 1967. *Intonation and Grammar in British English*. The Hague: Mouton.

Hauser, Marc, Chomsky, Noam, and Fitch, W. Tecumseh, 2002. "The Faculty of Language: What Is It, Who Has It, and How did it Evolve?" *Science* 298:1569–1579.

Ladd, D. Robert, 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

McWhinnie, Brian, 2005. “Item Based Constructions and the Logical Problem.” In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition. CoNNL-9*. New Brunswick: ACL, 53–68.

Neal, Radford and Hinton, Geoffrey, 1999. “A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants.” In Michael Jordan (ed.), *Learning in Graphical Models*, Cambridge, MA: MIT Press. 355–368.

Niyogi, Partha, 2006. *Computational Nature of Language Learning and Evolution*. Cambridge MA: MIT Press.

Pereira, Fernando and Schabes, Yves, 1992. “Inside-Outside Reestimation from Partially Bracketed Corpora.” In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*. ACL, 128–135.

Premack, David and Premack, Ann James, 1983. *The Mind of an Ape*. New York, NY: Norton.

Sakas, William and Fodor, Janet Dean, 2001. “The Structural Triggers Learner.”

In S. Bertolo (ed.), *Language Acquisition and Learnability*, Cambridge: Cambridge University Press. 172–233.

Siskind, Jeffrey, 1995. "Grounding Language in Perception." *Artificial Intelligence Review* 8:371–391.

Siskind, Jeffrey, 1996. "A Computational Study of Cross-Situational Techniques for Learning Word-to-Meaning Mappings." *Cognition* 61:39–91.

Steedman, Mark, 1996. "The Role of Prosody and Semantics in the Acquisition of Syntax." In James Morgan and Katherine Demuth (eds.), *Signal to Syntax*, Hillsdale, NJ: Erlbaum. 331–342.

Steedman, Mark, 2000. "Information Structure and the Syntax-Phonology Interface." *Linguistic Inquiry* 34:649–689.

Steedman, Mark, 2002a. "Formalizing Affordance." In *Proceedings of the 24th Annual Meeting of the Cognitive Science Society, Fairfax VA, August*. Mahwah NJ: Lawrence Erlbaum, 834–839.

Steedman, Mark, 2002b. “Plans, Affordances, and Combinatory Grammar.” *Linguistics and Philosophy* 25:723–753.

Steedman, Mark, 2007. “Compositional Semantics of Intonation.” *Submitted* .

Steels, Luc, 2004. “Constructivist Development of Grounded Construction Grammars.” In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*. Barcelona, 9–14.

Thornton, Rosalind and Tesan, Graciela, 2006. “Categorical Acquisition: Parameter Setting in Universal Grammar.” *Submitted* .

Tomasello, Michael, 1999. *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.

Villavicencio, Aline, 2002. *The Acquisition of a Unification-Based Generalised Categorial Grammar*. Ph.D. thesis, University of Cambridge.

Yang, Charles, 2002. *Knowledge and Learning in Natural Language*. Oxford: Oxford University Press.

Zettlemoyer, Luke and Collins, Michael, 2005. “Learning to Map Sentences to Logical Form: Structured Classification with Probabilistic Categorial Grammars.” In *Proceedings of the 21st Conference on Uncertainty in AI (UAI)*. ACL, 658–666.

Zettlemoyer, Luke S., Pasula, Hanna M., and Kaelbling, Leslie Pack, 2005. “Learning Planning Rules in Noisy Stochastic Worlds.” In *National Conference on Artificial Intelligence (AAAI),*. AAAI.