Computational Problem: Figure out the order of words (syntax)

- Jareth juggles crystals
  - Subject: Jareth
  - Verb: juggles
  - Object: crystals

Remember:
Children only see the output of the system (the observable word order of Subject Verb Object) and have to reverse engineer the generative process behind it.

Thinking About Syntactic Variation

- English: Subject Verb Object
- German: Subject Verb Object
- Kannada: Subject Verb Object

Similarities & Differences: Parameters

Chomsky: Different combinations of different basic elements (parameters) would yield the observable languages (similar to the way different combinations of different basic elements in chemistry yield many different-seeming substances).

Big idea: A relatively small number of syntax parameters yields a large number of different languages' syntactic systems.
Chomsky: Different combinations of different basic elements (parameters) would yield the observable languages (similar to the way different combinations of different basic elements in chemistry yield many different-seeming substances).

**Big Idea:** A relatively small number of syntax parameters yields a large number of different languages’ syntactic systems.

- **Total languages that can be represented**
  \[ 2^5 = 32 \]

---

**Similarities & Differences: Parameters**

Chomsky: Different combinations of different basic elements (parameters) would yield the observable languages (similar to the way different combinations of different basic elements in chemistry yield many different-seeming substances).

**Big Idea:** A relatively small number of syntax parameters yields a large number of different languages’ syntactic systems.

**Languages:**
- English
- French
- Japanese
- Navajo
- Tagalog

...
Chomsky: Children are born knowing the parameters of variation (and also potentially what values that can have). This is part of Universal Grammar. Input from the native linguistic environment determines what values these parameters should have.
Yang (2004): Learning Complex Systems Like Language

Only humans seem able to learn human languages. Something in our biology must allow us to do this.

This is what Universal Grammar is: innate biases for learning language that are available to humans because of our biological makeup (specifically, the biology of our brains).

But obviously language is learned, so children can’t know everything beforehand. How does this fit with the idea of innate biases/knowledge?

Observation: we see constrained variation across languages in their sounds, words, and structure. The knowledge of the ways in which languages vary is children’s innate knowledge.

Children know parameters of language variation… which they use to learn their native language.

The big point: even if children have innate knowledge of language structure, we still need to understand how they learn what the correct structural properties are for their particular language. One idea is to remember that children are good at tracking statistical information (like transitional probabilities) in the language data they hear.

The linguist-psychologist breakdown

Linguists
Charaterize “scope and limits of innate principles of Universal Grammar that govern the world’s languages”.

Psychologists
Emphasize the “role of experience and the child’s domain-general learning ability”.

Noam Chomsky
David Lightfoot
Michael Tomasello
Elizabeth Bates
Brian MacWhinney
Stephen Crain

Yang (2004): Learning Complex Systems
Yang (2004): Learning Complex Systems
Statistics for word segmentation (remember Gambell & Yang 2006)

"Modeling shows that the statistical learning (Saffran et al. 1996) does not reliably segment words such as those in child-directed English. Specifically, precision is 41.6%, recall is 23.3%. In other words, about 60% of words postulated by the statistical learner are not English words, and almost 80% of actual English words are not extracted. This is so even under favorable learning conditions".

Unconstrained (simple) statistics: not so good.

If statistical measure is constrained by language-specific knowledge (words have only one main stress), performance increases dramatically: 73.5% precision, 71.2% recall.

Constrained statistics - much better!

Yang (2004): Learning Complex Systems
Combining statistics with Universal Grammar

A big deal:
"Although infants seem to keep track of statistical information, any conclusion drawn from such findings must presuppose that children know what kind of statistical information to keep track of."

Ex: Transitional Probability
...of rhyming syllables?
...of syllables with nasal consonants?
...of syllables of the form CV (ba, ti)?

Linguistic Knowledge for Learning Structure

Parameters = constraints on language variation. Only certain rules/patterns are possible. This is linguistic knowledge.

A language’s grammar
= combination of language rules
= combination of parameter values

Idea: use statistical learning to learn which value (for each parameter) that the native language uses for its grammar. This is a combination of using linguistic knowledge & statistical learning.

Yang (2004): Variational Learning

Idea taken from evolutionary biology: In a population, individuals compete against each other. The fittest individuals survive while the others die out.

How do we translate this to learning language structure?
Yang (2004): Variational Learning

Idea taken from evolutionary biology:
A child’s mind consists of a population of grammars that are competing to analyze the data in the child’s native language.

Population of Grammars

Variational Learning Details

At any point in time, a grammar in the population will have a probability associated with it. This represents the child’s belief that this grammar is the correct grammar for the native language.

Native language data point

This grammar can analyze the data point while the other two can’t.
Variational Learning Details

Before the child has encountered any native language data, all grammars are equally likely. So, initially all grammars have the same probability, which is 1 divided the number of grammars available.

\[
\begin{align*}
\text{Prob} &= \frac{1}{3} \\
\text{Prob} &= \frac{1}{3} \\
\text{Prob} &= \frac{1}{3}
\end{align*}
\]

If there are 3 grammars, the initial probability for any given grammar = \(\frac{1}{3}\)

Variational Learning Details

As the child encounters data from the native language, some of the grammars will be more fit because they are better able to account for the structural properties in the data.

\[
\begin{align*}
\frac{1}{3} &\rightarrow \frac{4}{5} \\
\frac{1}{3} &\rightarrow \frac{1}{20} \\
\frac{1}{3} &\rightarrow \frac{3}{20}
\end{align*}
\]

Other grammars will be less fit because they cannot account for some of the data encountered. Grammars that are more compatible with the native language data will have their probabilities increased while grammars that are less compatible will have their probabilities decreased over time.

Variational Learning Details

After the child has encountered enough data from the native language, the native language grammar should have a probability near 1.0 while the other grammars have a probability near 0.0.

\[
\begin{align*}
\text{Prob} &= 1.0 \\
\text{Prob} &= 0.0 \\
\text{Prob} &= 0.0
\end{align*}
\]

Variational Learning Details

How do we know if a grammar can successfully analyze a data point or not?

Example: Suppose \(\heartsuit\) is the subject-drop parameter.

\(\heartsuit\) is +subject-drop, which means the language may optionally choose to leave out the subject of the sentence, like in Spanish.

\(\heartsuit\) is -subject-drop, which means the language must always have a subject in a sentence, like English.

Here, one grammar is +subject-drop while two grammars are -subject-drop.
Variational Learning Details

How do we know if a grammar can successfully analyze a data point or not?

Example data: \textit{Vamos = coming-1st-pl = “We’re coming”}

- The \textit{+subject-drop} grammar is able to analyze this data point as the speaker optionally dropping the subject.

\begin{itemize}
\item Prob = 1/3
\end{itemize}

- The \textit{-subject-drop} grammars cannot analyze this data point since they require sentences to have a subject.

\begin{itemize}
\item Prob = 1/3
\end{itemize}

Variational Learning Details

How do we know if a grammar can successfully analyze a data point or not?

Example data: \textit{Vamos = coming-1st-pl = “We’re coming”}

- The \textit{+subject-drop} grammar would have its probability increased if it tried to analyze the data point.

\begin{itemize}
\item Prob = 1/3
\end{itemize}

- The \textit{-subject-drop} grammars would have their probabilities decreased if either of them tried to analyze the data point.

\begin{itemize}
\item Prob = 1/3
\end{itemize}

Important idea: From the perspective of the subject-drop parameter, certain data will only be compatible with \textit{+subject-drop} grammars. These data will always reward grammars with \textit{+subject-drop} and always punish grammars with \textit{-subject-drop}.

\begin{itemize}
\item Certain data always reward \textit{+subject-drop} grammar(s).
\item Certain data always punish \textit{-subject-drop} grammar(s).
\end{itemize}

These are called \textit{unambiguous data} for the \textit{+subject-drop} parameter value because they unambiguously indicate which parameter value is correct (here: \textit{+subject-drop}) for the native language.

The Power of Unambiguous Data

Unambiguous data from the native language can only be analyzed by grammars that use the native language’s parameter value.

This makes unambiguous data very influential data for the child to encounter, since it is incompatible with the parameter value that is incorrect for the native language.

Ex: the \textit{-subject-drop} parameter value is not compatible with sentences that drop the subject. So, these sentences are unambiguous data for the \textit{+subject-drop} parameter value.

Important to remember: To use the information in these data, the child must know the subject-drop parameter exists.
Yang (2004): Learning Complex Systems

Learning Parametric Systems: Variational Learning

Grammars compete against each other to see which can best analyze the available data.

Added perk: Learning is then gradual (probabilistic).

Problem: Do unambiguous data exist for entire grammars? This requires data that are incompatible with every other possible parameter value of every other possible grammar.

This seems unlikely for real language data because parameters connect with different types of patterns, which may have nothing to do with each other.

Unambiguous issues

Parameter 1: subject-drop

Spanish: +subject-drop

Patterns allowed:
- Vamos
  - go-1st-pl-pres
  - “We go”

- Nosotros vamos
  - 1st-pl go-1st-pl-pres
  - “We go”

Parameter 2: Head-directionality

Edo/English: Head first

Basic word order:
- Subject Verb Object [SVO]

Prepositions:
- Preposition Noun Phrase
- Possessed before Possessor
- Possession Possessor

Preposition Noun Phrase
- Preposition
- NP
- Object
Parameter 2: Head-directionality

Japanese/Navajo: Head-final

Basic word order:
Subject Object Verb [SOV]

Postpositions:
Noun Phrase Postposition

Possessor before Possessed
Possessor Possession

Data point:
Subject Object Verb

Unambiguous issues

Grammars available:
+subj-drop +subj-drop
+head-first -head-first
-subj-drop -subj-drop
+head-first -head-first

G1 G2
G3 G4

G1?
+subj-drop allows Subject to be spoken
+head-first predicts SVO

Which grammars can analyze this data point?

G1 G2
+subj-drop +subj-drop
+head-first -head-first
-subj-drop -subj-drop
+head-first -head-first

G1 G2
+subj-drop +subj-drop
+head-first -head-first
-subj-drop -subj-drop
+head-first -head-first
There’s more than one grammar compatible with this data point... even though we feel like it should be informative for head directionality.
Yang (2004): Learning Complex Systems

Learning Parametric Systems: Variational Learning
Grammars compete against each other to see which can best analyze the available data.

Parameterized Grammars
This algorithm can take advantage of the fact that grammars are really sets of parameter values.
Parameter values can be probabilistically accessed.

The Learning Algorithm
For each data point encountered in the input
Choose a grammar probabilistically from available grammars by probabilistically accessing the parameter values.
If this grammar can analyze the data point, increase the probability of all participating parameters values slightly (reward)
Else decrease the probability of all participating parameters values slightly (punish)

Probability calculations:

- Prob = 0.2 * 0.3 * 0.2 * 0.3 * 0.1
- Prob = 0.8 * 0.7 * 0.2 * 0.7 * 0.1
- Prob = 0.2 * 0.7 * 0.2 * 0.7 * 0.9

Successful analysis

Unsuccessful analysis
Yang (2004): Learning Complex Systems
Learning Parametric Systems: Variational Learning
Grammars compete against each other to see which can best analyze the available data.

Problem ameliorated: unambiguous data much more likely to exist for individual parameter values instead of entire grammars.

Yang (2004): Learning Complex Systems
Variational Learning: Sample Case
Null subjects:
Parameter 1: Pro-drop, rely on unambiguous subject-verb agreement
Ex: Spanish, Italian (+pro-drop) Ex: English (-pro-drop)

✓ Yo puedo cantar.
   I: can-1st-sg sing-inf
   ‘I can sing’

✓ Puedo cantar.
   I: can-1st-sg sing-inf
   ‘I can sing’

✓ Hay lluvia.
   Is-3rd-sg rain
   ‘There is rain’

✓ There is rain.

Null subjects: 2 binary parameters, 4 grammars

+pro-drop, +topic-drop          -pro-drop, -topic-drop
Warlpiri, American Sign Language

+pro-drop, +topic-drop          -pro-drop, -topic-drop
Chinese English

What happens for an English-learning child?
In this study, Yang (2004) examines the acquisition of complex linguistic systems in child language development. Specifically, the focus is on the acquisition of pro-drop and topic-drop grammars across different languages such as Warlpiri, American Sign Language, Italian, Spanish, Chinese, and English.

**Null subjects:**
- 2 binary parameters, 4 grammars

**Variational Learning: Sample Case**

**What happens for an English-learning child?**

Pro-drop languages usually depend on rich subject-verb agreement morphology. English doesn’t have that, which is something a child will easily notice. Knock out +pro-drop grammars.

But this still leaves the +topic-drop option. What data will rule that out?

Answer: Expletive subjects. (Can’t topic-drop them.)

"There’s a goblin in the castle."  "It’s raining outside."  But this only occurs in 1.2% of the data (fairly rare).

**Variational Learning: General Predictions**

The time course of when a parameter is set depends on how frequent the necessary evidence is in child-directed speech.

Parameters set early: more unambiguous data
Parameters set late: less unambiguous data
Parameters set at the same time: equal quantity of unambiguous data

(more recent) table from Yang (2010), Comp Models of Syn Acq
**Additional Evidence for the importance of (un)ambiguity**

Hadley, Rispoli, Fitzgerald, & Bahnsen (2010): input informativity (how much ambiguity in the input) is the most consistent predictor for morphosyntactic growth.

Pelham (2011): input ambiguity affects how children acquire pronoun forms (“It appears children may be sensitive to levels of ambiguity such that low ambiguity may aid error-free acquisition, while high ambiguity may blind children to case distinctions, resulting in errors.”)

**Another case study for variational learning**

Explain why children’s early output consistently contains “optional infinitives” (OIs) that are ungrammatical in the adult language. They produce these incorrect forms at the same time that they produce correct “finite” forms.

<table>
<thead>
<tr>
<th>English</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct:</td>
<td>“Mummy goes to work.”</td>
</tr>
<tr>
<td>Occasional output:</td>
<td>“Mummy go to work”</td>
</tr>
</tbody>
</table>

**One explanation: Variational Learning Model**

Legate & Yang (2007)

Grammar options: +Tense (English) vs. –Tense (Mandarin Chinese)

OI errors results because initial hypothesis is –Tense. This lessens over time when unambiguous +Tense data are observed.

+Tense unambiguous data: Morphological marking

he goes home

Prediction:

Morphologically rich languages like Spanish have a very short OI stage because a large proportion of the input rewards +Tense (and punishes –Tense).

Morphologically poor languages like English have a longer OI stage because only a small proportion of the input rewards the [+Tense] grammar (and punishes –Tense).

### Another case study for variational learning

Note: Not just a matter of shortening the word form – sometimes, the incorrect form is actually longer (French, Dutch). Also, the word order sometimes changes (Dutch). This seems likely to be the result of some process happening in the child’s mind, rather than simple production error.

<table>
<thead>
<tr>
<th>French</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Input: “La poupée dort.”</td>
<td>The doll sleep-3rd-sg</td>
</tr>
<tr>
<td>Occasional output: “La poupée dormir”</td>
<td>The doll sleep-inf</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dutch</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Input: “Ik eet ijs.”</td>
<td>I eat-3rd-sg ice cream</td>
</tr>
<tr>
<td>Occasional output: “Ik ijs eten”</td>
<td>I ice cream eat-inf</td>
</tr>
</tbody>
</table>
One explanation: Variational Learning Model
Legate & Yang (2007)
Languages tested:
English, French, Spanish

Observed behavior seems to match unambiguous input distributions
OI duration:
- English (high) > French (moderately high) >> Spanish (very low)

+Tense unambiguous data:
- English > French >> Spanish

Possible critique (from Freudenthal et al. 2010)
Too easy because rates of OI are very different. What about Dutch and German, who have OI rates that are moderately high?

Another explanation: MOSAIC model
Freudenthal et al. (2010)
Model of Syntax Acquisition in Children: "MOSAIC is a constructivist model of language learning, with no built-in knowledge of syntactic categories or rules, which is implemented as a working computational model." – Algorithmic level?

"MOSAIC takes as input corpora of child-directed speech and learns to produce as output ‘child-like’ utterances that become progressively longer as learning proceeds...input corpora are fed through the model multiple times."

Input:
- "He will"
- "He wants"
- "Go home"
- "Go away"

Another explanation: MOSAIC model
Freudenthal et al. (2010)
- has a strong utterance-final bias in learning
  "MOSAIC does not encode a word or phrase unless everything that follows that phrase has already been encoded in the network."

- has a weak utterance-initial bias in learning
  "The utterance-initial bias enables MOSAIC to associate utterance-initial words and short (frequent) phrases with (longer) utterance-final phrases."

- represents declaratives and questions separately
  (so no underlying linkage between these forms)

Who could you see? has no relation to You could see him.

Another explanation: MOSAIC model
Freudenthal et al. (2010)
Where OI errors come from: Compound finites

English:
- He can go home.
Another explanation: MOSAIC model
Freudenthal et al. (2010)
Where OI errors come from: Compound finites

English:

- "He can go home." → "Go home"
- "He go home" (utterance-final bias)
- + weak utterance-initial bias + linking

Dutch (+ changed word order):

- Hij wil ijs eten.
- "He wants ice cream.
- "He wants to eat ice cream."

Dutch (+ changed word order):

- Hij wil ijs eten.
- "He wants ice cream.
- "He wants to eat ice cream."

utterance-final bias
Another explanation: MOSAIC model

Freudenthal et al. (2010)

Where OI errors come from: Compound finites

English:

He can go home. → "Go home", "He go home"
utterance-final bias
+ weak utterance-initial bias + linking

Dutch (+ changed word order):
Hij    ijz            eten. → "ijz eten", "Hij ijz eten"
He wants ice cream      eat-inf
"He wants to eat ice cream."
utterance-final bias
+ weak utterance-initial bias + linking

Freudenthal et al. (2010) Concluding Thoughts

"...it is clear that both the VLM and MOSAIC do a relatively good job of predicting the cross-linguistic data...if we focus on the results of the second set of analyses, it is clear that there are important lexical effects on the distribution of OI errors in children's speech that are difficult for the VLM to explain..."

"...A more lexically oriented input-driven account could probably deal with this problem relatively easily by simply distinguishing between what the child is learning about copulas and auxiliaries and what the child is learning about lexical verbs, and predicting high levels of OI errors on lexical verbs and lower levels of OI errors on copulas and auxiliaries. Interestingly, this is exactly the pattern of results reported in two recent lexically oriented analyses of early child English (Wilson, 2003; Pine, Conti-Ramsden, Joseph, Lieven & Serratrice, 2008)."

Another complex system: Metrical phonology

Observable data: stress contour

EMphasis

( S S ) S
( H L ) H
EM pha sis

( S S S )
EM pha sis

Another complex system: Metrical phonology

Linguistic parameters = finite (if large)
hypothesis space of possible grammars

Which syllable of a larger unit is stressed?
{Leftmost, Rightmost}

Are all syllables included?
{Yes, No-not leftmost, No-not rightmost}

Are syllables differentiated?
{No, Yes-2 distinctions, Yes-3 distinctions}
Another complex system: Metrical phonology

**Comparing knowledge representations**
One traditional motivation for proposals of knowledge representation (such as parameters): The knowledge representation helps explain the constrained variation observed in adult linguistic knowledge across the languages of the world.

Another (sometimes implicit) motivation for proposals of knowledge representation: Having this knowledge representation pre-specified allows children to acquire the right generalizations from the data as quickly as they seem to do.

Pearl 2008, Pearl 2009, Pearl 2011: English metrical phonology
- Using computational methods and available empirical data, we can quantify this argument and explicitly test different proposals for knowledge representation
- At the same time, we can explore how acquisition could proceed if children were using these different knowledge representations

Another complex system: Metrical phonology

**English metrical phonology: Legate & Yang 2011**

English stress is a volatile area theoretically (a hard system to capture because of all the exceptions); developmental work could shed light on the target state for adults.

Some empirical data on children’s stress knowledge is available

Because of the need to capture both “core” and “exceptional” data, the system is very interesting from a developmental point of view.

Comparison of two proposals for metrical phonology systems, using a quantitative definition of productivity.

Idea: Productive rule system is one that will be adopted.