Bayesian Updating in Human Language Learning

Lisa Pearl

Oct 24, 2006

Road Map

- Introduction
  - Bayesian Updating Overview
  - Human Language Learning Overview
  - Mapping Between
- Case Studies
  - Syntax/Semantics
  - Syntax
  - Metrical Phonology

Road Map

- Introduction
  - Bayesian Updating Overview
  - Human Language Learning Overview
  - Mapping Between
- Case Studies
  - Syntax/Semantics
  - Syntax
  - Metrical Phonology

Introduction: Bayesian Updating

- Used to estimate the probability of a number of hypotheses, based on input
- The hypothesis space can be set up in a number of ways, which affects how the input distribution alters the probabilities

Bayesian Updating: Hypothesis Spaces

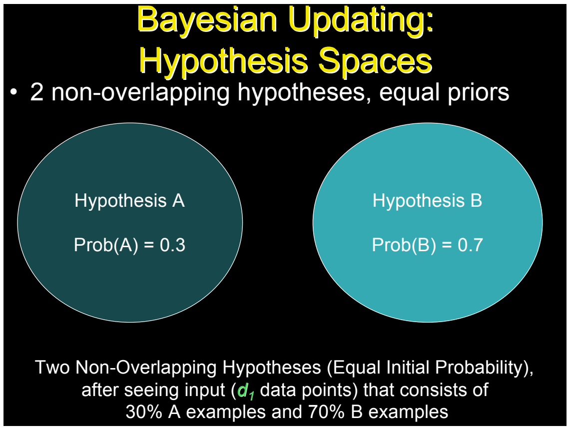
- 2 non-overlapping hypotheses, equal priors

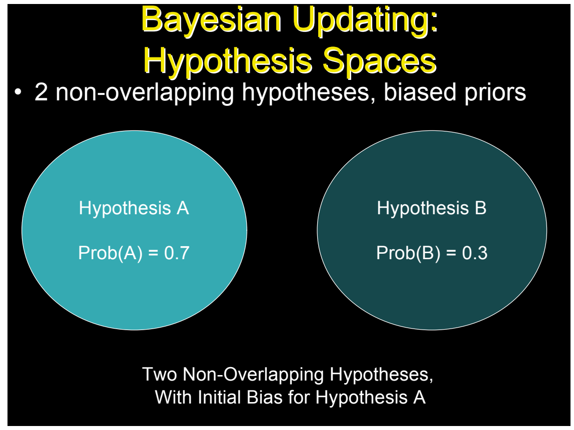
Two Non-Overlapping Hypotheses,
Equally Probable Initially

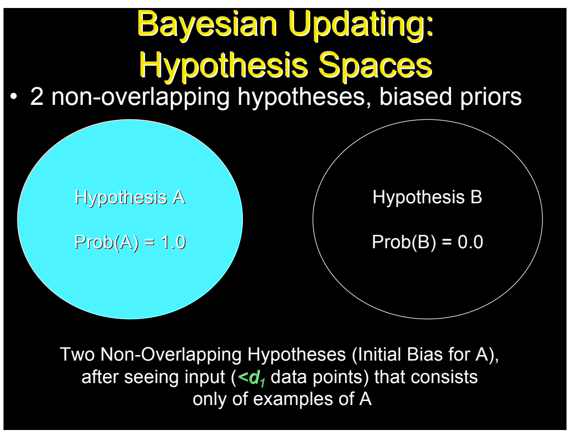
Bayesian Updating: Hypothesis Spaces

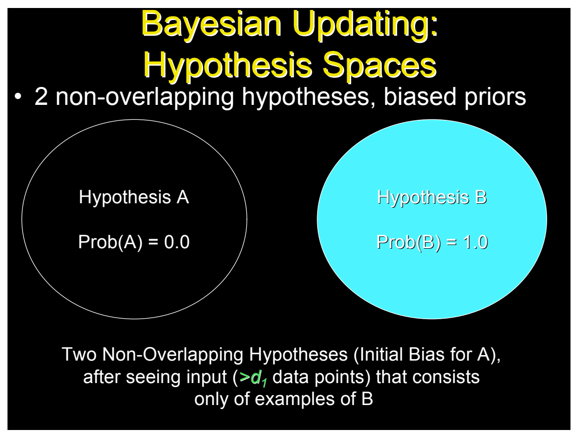
- 2 non-overlapping hypotheses, equal priors

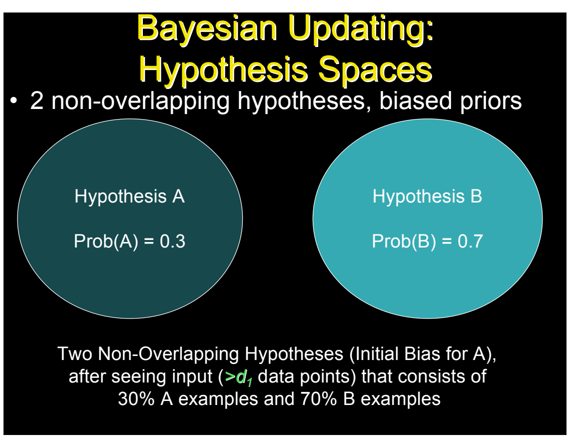
Two Non-Overlapping Hypotheses (Equal Initial Probability),
after seeing input (*d_1* data points) that consists
only of examples of A

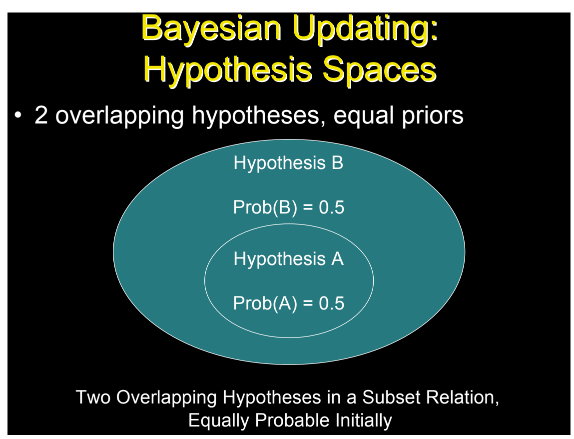

Bayesian Updating:
Hypothesis Spaces
• 2 non-overlapping hypotheses, equal priors
Hypothesis A
Prob(A) = 0.3
Hypothesis B
Prob(B) = 0.7
Two Non-Overlapping Hypotheses (Equal Initial Probability),
after seeing input (d_1 data points) that consists of
30% A examples and 70% B examples


Bayesian Updating:
Hypothesis Spaces
• 2 non-overlapping hypotheses, biased priors
Hypothesis A
Prob(A) = 0.7
Hypothesis B
Prob(B) = 0.3
Two Non-Overlapping Hypotheses,
With Initial Bias for Hypothesis A


Bayesian Updating:
Hypothesis Spaces
• 2 non-overlapping hypotheses, biased priors
Hypothesis A
Prob(A) = 1.0
Hypothesis B
Prob(B) = 0.0
Two Non-Overlapping Hypotheses (Initial Bias for A),
after seeing input ($<d_1$ data points) that consists
only of examples of A

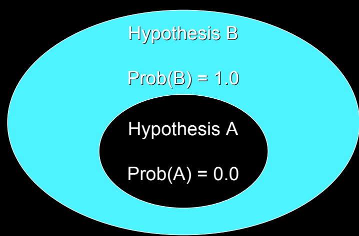

Bayesian Updating:
Hypothesis Spaces
• 2 non-overlapping hypotheses, biased priors
Hypothesis A
Prob(A) = 0.0
Hypothesis B
Prob(B) = 1.0
Two Non-Overlapping Hypotheses (Initial Bias for A),
after seeing input ($>d_1$ data points) that consists
only of examples of B


Bayesian Updating:
Hypothesis Spaces
• 2 non-overlapping hypotheses, biased priors
Hypothesis A
Prob(A) = 0.3
Hypothesis B
Prob(B) = 0.7
Two Non-Overlapping Hypotheses (Initial Bias for A),
after seeing input ($>d_1$ data points) that consists of
30% A examples and 70% B examples


Bayesian Updating:
Hypothesis Spaces
• 2 overlapping hypotheses, equal priors
Hypothesis B
Prob(B) = 0.5
Hypothesis A
Prob(A) = 0.5
Two Overlapping Hypotheses in a Subset Relation,
Equally Probable Initially

Bayesian Updating: Hypothesis Spaces

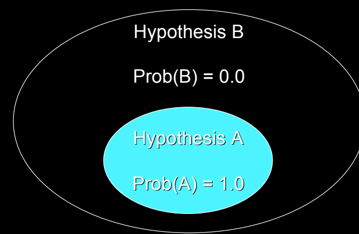
- 2 overlapping hypotheses, equal priors



Two Overlapping Hypotheses in a Subset Relation, after seeing input (d_2 data points) that consists only of examples of B

Bayesian Updating: Hypothesis Spaces

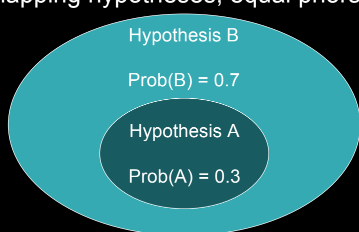
- 2 overlapping hypotheses, equal priors



Two Overlapping Hypotheses in a Subset Relation, after seeing input ($> d_2$ data points) that consists only of examples of A

Bayesian Updating: Hypothesis Spaces

- 2 overlapping hypotheses, equal priors



Two Overlapping Hypotheses in a Subset Relation, after seeing input ($> d_2$ data points) that consists of 30% A examples and 70% B examples

Bayesian Updating

- Bayesian updating is a domain-general updating procedure that can be integrated with other components of a learning theory that are domain-specific

Road Map

- Introduction
  - Bayesian Updating Overview
  - Human Language Learning Overview
  - Mapping Between
- Case Studies
  - Syntax/Semantics
  - Syntax
  - Metrical Phonology

Human Language Learning: Domain-General vs. Domain-Specific

- Examples of cognitive domains: vision, geometric representation, language
- Domain-general: not associated with any particular domain - can be used within any domain and across domains
- Domain-specific: associated with a particular domain - only used within this domain

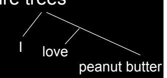
Human Language Learning

- Learning theory components
  - Representations of knowledge
  - Filters on data used as *intake* by learner
  - Procedure to update probability of different hypotheses, based on *intake*

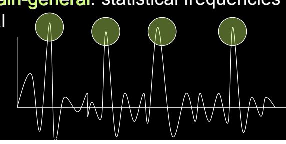
Human Language Learning

- Learning theory components for language
  - Representations of knowledge
    - Domain-specific: linguistic representations such as phonemes, morphemes, phrase structure trees

p^h…p…b^h…b… peanut+butter



    - Domain-general: statistical frequencies in the acoustic signal



Human Language Learning

- Learning theory components
  - Filters on data used as *intake* by learner
    - Domain-specific: use only main clause data (Lightfoot, 1991)

Rarely do I think that passing up peanut butter is a good idea.

    - Domain-general: use as much data as will fit in working memory at one time
      [ex: 7 words at a time]

Rarely do I think that passing up peanut butter is a good idea.

Human Language Learning

- Learning theory components
  - Procedure to update probability of different hypotheses, based on *intake*
    - Domain-specific: Trigger Learning Algorithm (Gibson & Wexler, 1994)
    - Domain-general: Bayesian Updating

Road Map

- Introduction
  - Bayesian Updating Overview
  - Human Language Learning Overview
  - Mapping Between
- Case Studies
  - Syntax/Semantics
  - Syntax
  - Metrical Phonology

Mapping Between

- human language learning:
  - What children know: knowledge of language
    - Can discover this from theoretical linguistics work
  - When children know it: trajectory of knowledge acquisition
    - Can discover this from experimental linguistics work
  - *How* do children learn it: the process that causes children to acquire the appropriate "what" by the appropriate "when"
    - Can explore this with computational modeling work

Exploring the "How" of Human Language Learning

- Assumptions:
  - Have domain-specific representations of knowledge available (hypotheses about the adult language)
  - Learner's task: determine the probabilities of the various hypotheses available
  - Learner uses domain-general procedure of Bayesian updating to shift probability between the various hypotheses, based on the intake

Exploring the "How" of Human Language Learning

- Is this enough, or does the learner need some kind of filter on the available input so that the learner's *intake* consists of some subset of the input? If filters are required, what sort are they?
- *Let's look at some case studies in human language learning and find out…*

Road Map

- Introduction
  - Bayesian Updating Overview
  - Human Language Learning Overview
  - Mapping Between
- Case Studies
  - Syntax/Semantics
  - Syntax
  - Metrical Phonology

Syntax/Semantics: Anaphoric *One*

- Knowledge (the "what"):

"Jack has a red ball, and Lily has *one*, too."

Adult intuition check:
What color ball does Lily have?

Syntax/Semantics: Anaphoric *One*

- Knowledge (the "what"):

"Jack has a red ball, and Lily has *one*, too."

Adult intuition check:
What color ball does Lily have?

(usually) a red ball

Syntax/Semantics: Anaphoric *One*

- Knowledge (the "what"):

"Jack has a red ball, and Lily has *one*, too."

Syntax (structure):
one has "*red ball*" as its linguistic antecedent (*one* is anaphoric to "*red ball*")

Syntax/Semantics: Anaphoric *One*

- Knowledge (the "what"):

?

"Jack has a red ball, and Lily has *one*, too."

Semantics (meaning):

the referent of *one* has the property mentioned in the linguistic antecedent of *one* (*red*)

Syntax/Semantics: Anaphoric *One*

- Knowledge (the "what"):

"Jack has a red ball, and Lily has *one*, too."

Semantics (meaning):

the referent of *one* has the property mentioned in the linguistic antecedent of *one* (*red*)

Syntax/Semantics: Anaphoric *One*

But what other possibilities are there?

Syntax/Semantics: Anaphoric *One*

- Knowledge (the "what"):

"Jack has a red ball, and Lily has *one*, too."

Syntax (structure) - other possibility:

one has "*ball*" as its linguistic antecedent (*one* is anaphoric to "*ball*")

Syntax/Semantics: Anaphoric *One*

- Knowledge (the "what"):

?

"Jack has a red ball, and Lily has *one*, too."

Semantics (meaning) - other possibility:

the referent of *one* has no restriction on its property (any property is acceptable)

Syntax/Semantics: Anaphoric *One*

- Knowledge (the "what"):

"Jack has a red ball, and Lily has *one*, too."

Semantics (meaning) - other possibility:

the referent of *one* has no restriction on its property (any property is acceptable)

Syntax/Semantics: Anaphoric *One*

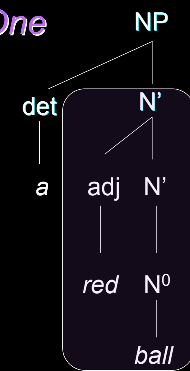
Nonetheless, adults do not favor this second interpretation. So, children must learn that the first interpretation is the correct one. What does their hypothesis space look like?

Syntax/Semantics: Anaphoric *One*

Syntactic Structure

"Jack has a red ball, and Lily has ***one***, too."

one refers to the N' "*red ball*"

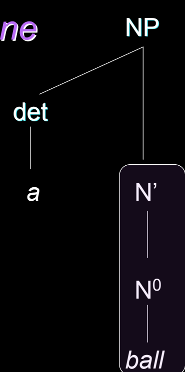


Syntax/Semantics: Anaphoric *One*

Syntactic Structure

"Jack has a ball, and Lily has ***one***, too."

one refers to the N' "*ball*"

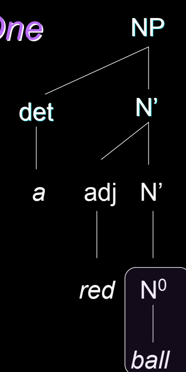


Syntax/Semantics: Anaphoric *One*

Syntactic Structure

"Jack has a red ball, and Lily has ***one***, too."

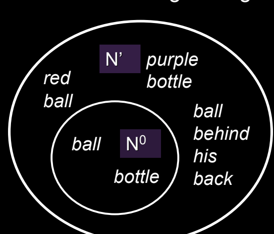
one refers to the N^0 "*ball*"



Syntax/Semantics: Anaphoric *One*

Syntactic Hypothesis Space (Subset-Superset Relation)

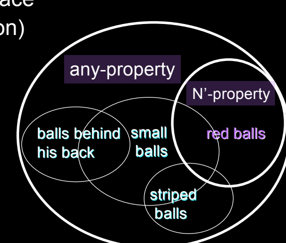
"***one*** has as its antecedent strings categorized as…"



Syntax/Semantics: Anaphoric *One*

Semantic Hypothesis Space (Subset-Superset Relation)

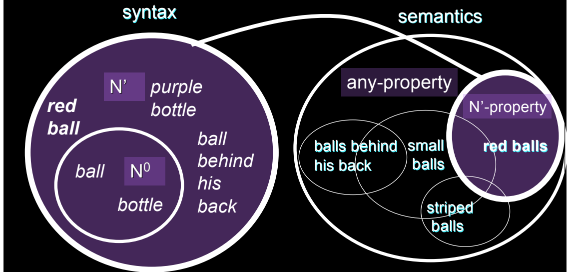
"the referent of ***one*** refers to objects that are…"



"Jack has a red ball and Lily has *one*, too."

Syntax/Semantics: Anaphoric *One*

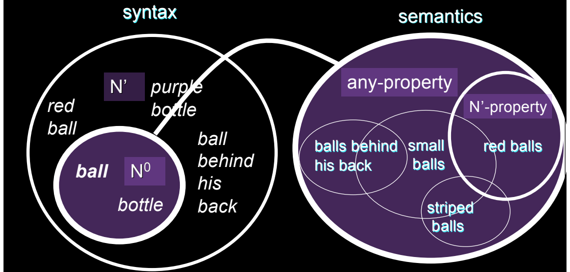
- Link between the two linguistic domains



"…*red ball*…*one*…"

Syntax/Semantics: Anaphoric *One*

- Link between the two linguistic domains



"…*ball*…*one*…"

Syntax/Semantics: Anaphoric *One*

- The "when" of anaphoric *one*:

Lidz, Waxman, & Freedman (2003) demonstrated experimentally that 18-month old children behave as if they have the adult knowledge:

- *one* has an antecedent that is N' ("*red ball*")
- the referent of *one* has the property mentioned in the N' antecedent (red)

Syntax/Semantics: Anaphoric *One*

So *how* do children converge on the correct hypotheses in these two (connected) domains?

Syntax/Semantics: Anaphoric *One*

- Lidz, Waxman, & Freedman (2003) analyzed the data available to children, and found that less than 0.3% of it is unambiguous evidence for the correct hypotheses
- Given this data sparseness, they concluded that children must either already have this knowledge (innate bias/domain-specific knowledge) or else derive it by other means

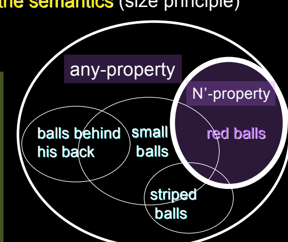
Syntax/Semantics: Anaphoric *One*

- Regier & Gahl (2004) replied that the domain-general procedure of Bayesian updating could converge on the correct answer because some of the ambiguous data could be used to converge on the subset in the semantics (size principle)

The referent of *one* has…

Size principle: if only data from the subset are encountered, the learner is increasingly biased to believe there is a restriction to the subset (Tenenbaum & Griffiths, 2001)

"Jack has a red ball and Lily has *one*, too."



Syntax/Semantics: Anaphoric *One*

Regier & Gahl's conclusion: a domain-general updating procedure is sufficient to converge on the correct knowledge of anaphoric *one* - no domain-specific biases required

Syntax/Semantics: Anaphoric *One*

- Pearl & Lidz (in prep) reply: Using only *some* of the available data is a bias (domain-specific filter).
- What happens if Bayesian updating is used for all the available data? This is the true test for how a domain-general updating procedure fares by itself.

Syntax/Semantics: Anaphoric *One*

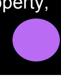
- The learner ends up with the wrong answer in *both* linguistic domains

"Jack has a red ball and Lily has *one*, too."

Bayesian Updating with all available data:
- Syntax: *one* refers to the N^0 *ball,* not the N' *red ball*

N^0
|
ball

- Semantics: *one* refers to a ball with any property, not the N'-property red



Syntax/Semantics: Anaphoric *One*

- This happens because a large portion of the available data, though ambiguous, still biases the learner towards the incorrect hypotheses in both the syntactic and semantic domain
- Conclusion: need a domain-specific filter to ignore a large portion of the ambiguous data (bias to use subset of the available data when using Bayesian updating)

Road Map

- Introduction
  - Bayesian Updating Overview
  - Human Language Learning Overview
  - Mapping Between
- Case Studies
  - Syntax/Semantics
  - Syntax
  - Metrical Phonology

Syntax: Old English Word Order

- Old English Word Order (YCOE, PPCME2)

1000 A.D. - 1150 A.D.: mostly Object Verb (OV) order

…Object Verb…

1200 A.D.: mostly Verb Object (VO) order

…Verb Object…

Syntax: Old English Word Order

- Old English Word Order (YCOE, PPCME2)

1000 A.D. - 1150 A.D.: mostly Object Verb (OV) order

he$_{\text{Subj}}$ Gode$_{\text{Obj}}$ þancode$_{\text{TensedVerb}}$
he God thanked
'He thanked God'
(*Beowulf, 625*)

1200 A.D.: mostly Verb Object (VO) order

& [mid his stefne]$_{\text{PP}}$ he$_{\text{Subj}}$ awecð$_{\text{TensedVerb}}$ deade$_{\text{Obj}}$ [to life]$_{\text{PP}}$
& with his stem he awakened the-dead to life
"And with his stem, he awakened the dead to life."
(*James the Greater*, 30.31)

Syntax: Old English Word Order

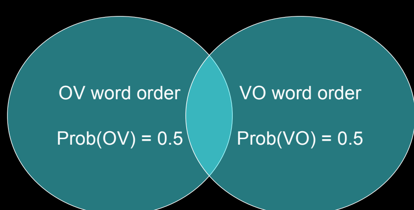
Adult Word Order Knowledge: probability distribution between the two word order options; changes over time

1000 A.D. - 1150 A.D.
Access OV order option ~77% of the time
Access VO order option ~23% of the time

1200 A.D.
Access OV order option ~25% of the time
Access VO order option ~75% of the time

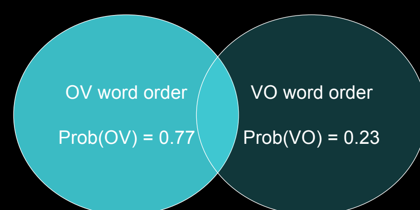
Syntax: Old English Word Order

- Hypothesis Space: Two overlapping hypotheses, equal priors



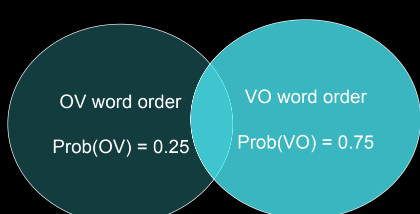
Syntax: Old English Word Order

- Correct adult probability distribution between 1000 A.D. and 1150 A.D.



Syntax: Old English Word Order

- Correct adult probability distribution at 1200 A.D.



Syntax: Old English Word Order

- So how does language change help us answer questions about language learning?
- Assumption (Lightfoot, 1991): For Old English, the population-level shift is due to individuals misconverging on the correct probability distribution, compounded over time.
- Individual misconvergence happens during learning

Syntax: Old English Word Order

- So how does language change help us answer questions about language learning?
- Simulate population of Old English speakers with individuals who use a particular learning mechanism (i.e. Bayesian updating, with or without filters on data intake)

Syntax: Old English Word Order

- So how does language change help us answer questions about language learning?
- Individuals at each point in time will misconverge on the probability distribution
- If the amount of individual misconvergence at each point in time is correct, the population as a whole will shift its probability distribution the correct amount at the correct times

Syntax: Old English Word Order

- So how does language change help us answer questions about language learning?

Logic:

(1) Population-level behavior is correct (language change)

(2) Population-level behavior is result of individual-level behavior

(3) Individual-level behavior is result of learning mechanism implemented

Assumption: learning mechanism is correct.

Syntax: Old English Word Order

Simulation Algorithm:

Create Old English population at time 1000 A.D.
Every 2 years until 1200 A.D.
oldest members die off
new members receive data from remaining population & use learning mechanism to converge on probability distribution between OV and VO word order

Syntax: Old English Word Order

- Objective:
  - 1000 A.D. - 1150 A.D.
  - OV = ~77%, VO = ~23%
  - 1200 A.D.
  - OV = 25%, VO = ~75%

Learning Mechanism in individuals:

- Bayesian updating by itself
- Bayesian updating with domain-specific filters

Syntax: Old English Word Order

- Bayesian Updating by itself (no filters on data intake):
  - The population does not behave correctly (too much probability is shifted to the VO option too soon)
  - Therefore, individuals not behaving correctly.
  - Therefore, not an accurate model of individual learning.

Syntax: Old English Word Order

- Bayesian Updating with domain-specific filters
  - Filter 1: use only data in main clauses

Jack told Lily that he had to go off on an epic adventure.

  - Filter 2 : use only data that is unambiguous
  - The population behaves correctly
  - Therefore, individuals behaving correctly.
  - Therefore, an accurate model of individual learning.

Syntax: Old English Word Order

- (Familiar) Conclusion: need domain-specific filters to ignore a large portion of the available data (bias to use subset of the available data when using Bayesian updating)

Road Map

- Introduction
  - Bayesian Updating Overview
  - Human Language Learning Overview
  - Mapping Between
- Case Studies
  - Syntax/Semantics
  - Syntax
  - Metrical Phonology

Metrical Phonology

- Metrical phonology is what tells us to put the

EMphasis on a certain SYLlable

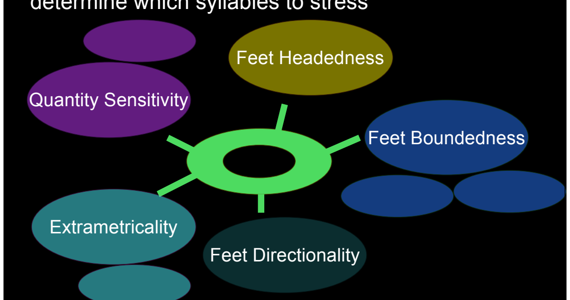
instead of putting the

emPHAsis on a different sylLAble

(emphasis often referred to as 'stress')

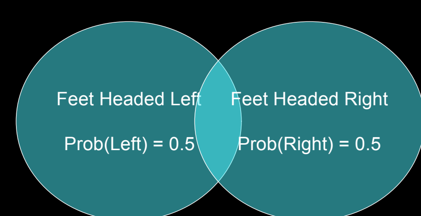
Metrical Phonology

- 5 main parameters and 4 sub-parameters that determine which syllables to stress



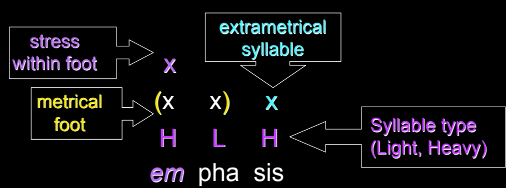
Metrical Phonology

- Each of the parameters is a hypothesis space that is overlapping



Metrical Phonology

- *All* the parameters interact with each other to produce the observable stress contour of a word



Metrical Phonology

- The learner must take the available data (observable stress contours) and determine which of the two hypotheses for each parameter is correct for a given data point.

- This is quite hard!

Input

em pha sis

Signifies

Feet Headed Left?
Quantity Sensitive?
Feet Direction Right?
Extrametricality?
…

Metrical Phonology

- Metrical Phonology Parameters for English:
  - Quantity Sensitive (classify syllables as Light/Heavy)
    - Syllables with consonants on the end ('em') are considered Heavy
  - Extrametricality (one syllable is not included in a metrical foot)
    - The rightmost syllable is not included in a metrical foot
  - Bounded Feet (a metrical foot is of a certain size)
    - 2 units make a foot, a syllable is a unit
  - Feet Headedness Left (stress falls on the leftmost syllable in a foot)
  - Feet Directionality Right (metrical feet are constructed right to left)

Metrical Phonology

- This is hard enough to learn, but English data makes it even harder. While there are data that implicate the correct hypotheses for English, there are also many *exceptions* that implicate the incorrect hypotheses for English.

- For example, English is a language that is Quantity Sensitive. Yet, there are data that can only be accounted for if the opposite value (Quantity Insensitive) is used.

Metrical Phonology

- While Bayesian Updating is again a sensible procedure to use for shifting probabilities between competing hypotheses, the trick is what the learner's data intake is.

- Feasibility study: Is it possible for a Bayesian learner to converge on the correct hypotheses for each of the 5 parameters and 4 sub-parameters in the metrical phonology system, given realistic English data?

Metrical Phonology

- Let's try a filter on data intake: use only data that is unambiguous (as perceived by the learner)

- This will again cut down on the data used, since the learner is only using a subset of the available data. Moreover, determining that a given data point is unambiguous for any of the 9 hypothesis spaces is no trivial feat.

Metrical Phonology

- But luckily, this works!
- Given data distributions estimated from ~500,000 words of child-directed speech, a Bayesian learner that uses only data it perceives as unambiguous can converge on the correct hypotheses for all the parameters of English

Metrical Phonology

- (Familiar) Conclusion: Bayesian updating succeeds when paired with domain-specific filters that ignore a large portion of the available data (bias to use subset of the available data when using Bayesian updating)

So what have we seen?

- Human language learning problems seem to require domain-specific filters on data intake in addition to a domain-general learning procedure such as Bayesian updating
  - Syntax/Semantics: Anaphoric *one*
    - Works only if it ignores some ambiguous data
  - Syntax: OV/VO Word Order
    - Works only if uses only main clause data & unambiguous data
  - Metrical Phonology: (Hard Case) English
    - Works if uses unambiguous data

The End