Statistical models of language acquisition

Micha Elsner (some material from Mark Johnson)

August 22, 2012

Me

- New faculty in Linguistics
- Ph.D. at Brown (advisor Eugene Charniak, worked a bit with Mark Johnson)
- Postdoc at U. of Edinburgh (supervisor Sharon Goldwater)
- My acquisition research: learning lexicon from data with phonetic variation
  - ...by learning a model of surface phonetics
- Part of a field-wide effort to model acquisition all the way from acoustics to words

Overview

Why

- Why child language acquisition?
- Why modeling?
- Why statistics?

What

- Developmental linguistics in one slide
- Some high-level questions
- Areas of interest

The course

- Course objectives
- Assessment

Child language acquisition

Process by which babies learn their first language(s):

- What evidence do they use?
- What sort of grammars do they build?
  - How do these change over time?
- What is the mechanism of learning?
  - Is it general or language-specific?

(Linguistics, cognitive science, developmental psychology etc...)

Parallel work in engineering

Learning about language with little data:

- Under-resourced languages
  - Ad placement on Polish eBay
- Endangered/extinct languages
  - Translating Hittite inscriptions
- Rapidly changing language
  - Hip-hop slang in product reviews
- What evidence can we get for free (or cheaply)?
- What sort of grammars give us good results?
- How do we learn *quickly, reliably, scalably*?

(Computer science, machine learning, statistics)

Motivations

- Cognitive and engineering motivations are quite similar...
- Often the same people care about both
- **However**, important to be honest about motivations...
- What resources can we give our algorithm?
  - Cognitive: proposed linguistic universals, cognitively plausible biases, analyses we believe the baby can already do
  - Engineering: a few labeled examples, lots of data in similar but better-resourced language, implicit evidence from punctuation/orthography

Why modeling?

Models

Formal descriptions of cognitive processes...

Descriptive:

- Describes the trend without making causal claims
- Child's vocabulary size as function of age

Explanatory:

- Implements a proposed mechanism
- Attempts to reproduce a known trend
- eg, knowing some words gives basis for learning others faster

This course: mostly explanatory models

What can we learn from modeling?

Having a model that “works” doesn’t mean we have the right model...

> Increasing f-score is often not a scientific contribution
> but *how you did it* may be a scientific contribution
> (Mark Johnson)

> “Animals don’t move on wheels” (Tom Wasow qtd. in Johnson)

What sort of claims can we support with models?

Claims we can make

- **Learnability**: X sources of evidence tell us about Z
  - Could falsify poverty-of-stimulus style claims (Z can't be learned from X)
- Does *failure* to learn tell us anything?
  - Yes, but... could always mean you set up the model wrong
- **Learning synergies**: helps to learn X and Y together
- **What to look for experimentally**: model predicts effects we can search for in real world
  - "Nobody thought the heart was a pump until there were pumps" — Eugene Charniak
  - Which instances are hardest? What are typical mistakes?
  - cf. syntactic surprisal effects on reading time
- **Potential mechanisms**: usually *not* enough to look at model results
  - Learning trajectory (pattern of mistakes over time)
  - Reaction times, neural measurements
  - Can be murky... easy to make specious claims
  - Neural nets $\neq$ biologically plausible computation

Marr's three levels

David Marr (psych. of vision):

- Computational (Johnson: informational)
  - What information does the system use? What is its objective?
- Algorithmic
  - What low-level representation is used? What computations are carried out?
- Physical/implementational
  - What does the hardware look like and what does it do?

Evaluating learnability generally a job for *computational-level* models... mechanisms perhaps at *algorithmic* level.

Performance

Easy to be insulting about performance...

- Would rather see better explanations than higher f-scores, etc

True to a degree... but:

- Performance keeps you honest
  - Can't mistakenly present a naive model with complicated decorations
- Your favorite evidence perhaps already covered by more general features
- Mistakes made by dumb models are less interesting

Why statistics?
Answer 1: a framework for **rational models**

- Rational models: study human cognitive processes in terms of their *objectives*
- Reaction against eg. Kahneman and Tversky's studies of cognitive biases
  - K+T: Humans are poor reasoners in the classical sense
  - Focus on *processes* of human reasoning
- Rationality: consider processes as attempts to *optimally/rationally/normatively* achieve some *goal*
  - Proponents: John Anderson, David Marr, Nick Chater etc
  - Goal may be general or very task-specific (cognitive biases often result from misusing a task-specific goal)
  - Some bias introduced if goal cannot be efficiently achieved (bounded rationality)
  - But evolution searches for good solutions for a given problem, so sensible to study optimal mechanism

Rationality

Rational cognition approach: statistics is good for describing cognitive objects and optimizing them

- Doesn't matter if this is how humans do it
- Focus on getting the objective right, not the process

Good things about statistics

- Good off-the-shelf learning mechanisms
- Mathematical theorems about optimal learning/estimation mechanisms

Why statistics?
Answer 2: humans use statistics

Saffran, Aslin and Newport 96

Infants familiarized with stream of syllables:

`bidakupadotigolabu...`

- Within-word syllable transitions deterministic: `bi` always followed by `da`
- Across boundaries, transitions non-deterministic: `ku` followed by `bi,pa,go`

After 2 minutes of training, infants can tell stream of words:

`golabugolabugolabu`

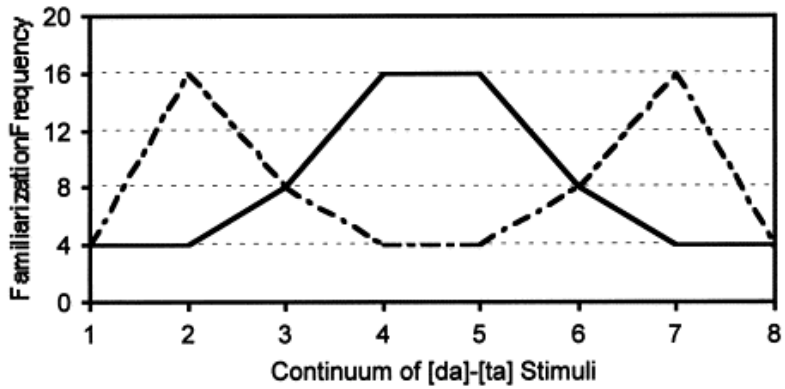
From nonwords:

`gobipagobipagobipa`

Humans use statistics (2)

Maye, Werker and Gerken 01

Infants familiarized with unimodal or bimodal distribution of speech sounds



- Infants in bimodal condition appear to form two categories
- In unimodal condition, they don't distinguish endpoints of continuum
- Argument: category formation driven by frequencies

Humans

Human-oriented approach: people are actually (approximately) doing statistical computations

- Some neurological work aimed at finding out how
- Focus on mechanisms for computing with bounded resources

Good things about statistics

- Most ways of getting Saffran-like results require some kind of statistical estimation
- Some plausible biological mechanisms proposed (eg Shi and Griffiths 09)
- Possible to be mathematically clear about how things are being approximated

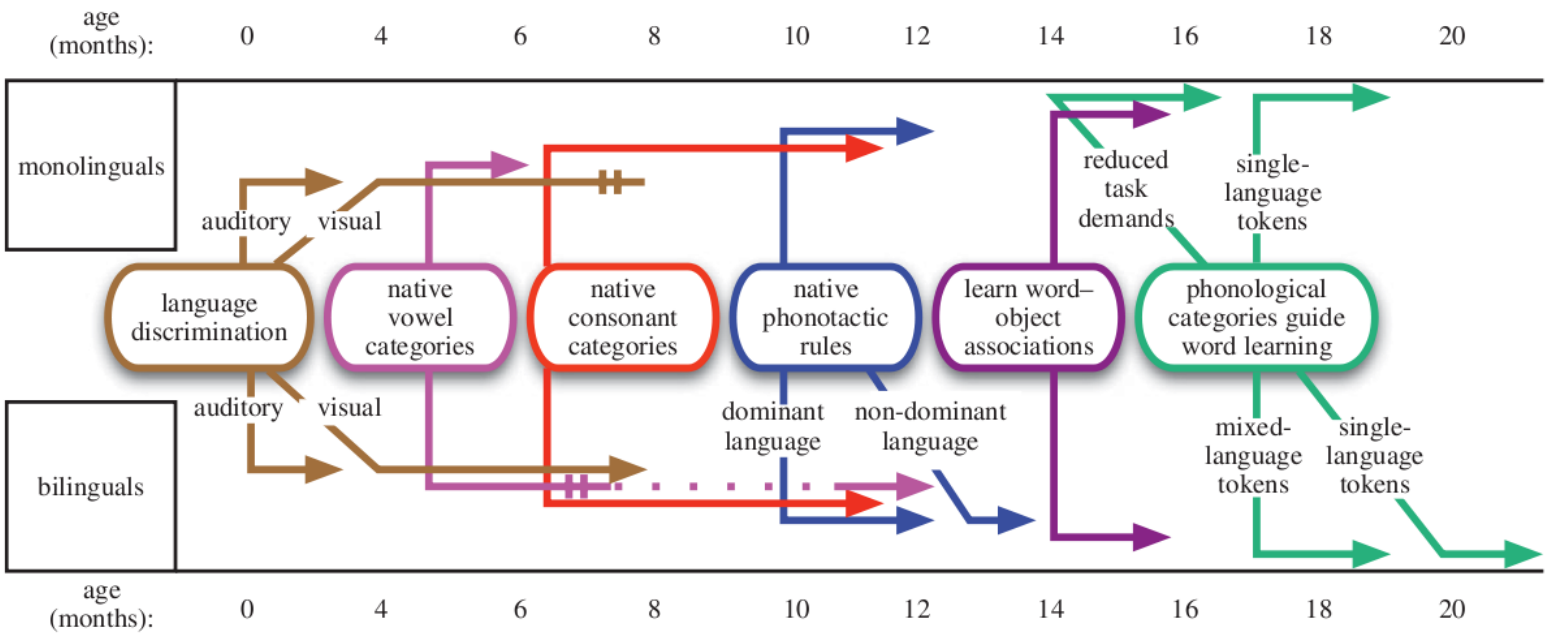
Why statistics? The "learn one toolkit" approach (Other methods often more similar than you think)

Many "hot, new" methods for building cognitive models. Takes a long time to learn them all– we'll focus mainly on **Bayesian** framework.
However, Bayesian models similar/generalize:

- Classical model selection (BIC, AIC)
- Some kinds of dimensionality reduction (like PCA)
- Minimum description length
- Many kinds of neural nets
- Connections to spectral algorithms in some cases

At *algorithmic* level, still matters which you use. At the *computational* level, may not.

Developmental linguistics in one slide



(Werker, Byers-Heinlein and Fennell 09)

Child language development is...

Robust

- Neurotypical humans all learn their native language...
  - (Up to variance in adult grammars; see Ewa Dabrowska and others)
- All features of any NL must be learnable
- Learning affected, but not stopped by exposure to multiple languages/dialects
- Or to speech errors/other naturally occurring ungrammaticality
- Child-directed speech not a must
- Little explicit instruction from parents
  - Chomsky's famous point about negative evidence

Child language development is...

Synergistic

- Many different things are learned at once
- But not *everything* is learned at once
- The traditional subfields of linguistics (incl. this course) should not mislead us about how the infant divides up the problem

Incremental

- Infant has limited memory
- Plausible mechanisms *don't* store all previous experiences
- But still capable of discarding previous hypotheses
- (Problematic for current computational methods)

Some larger controveries

Nativism (Universalism)

- Strong: Mechanisms of language learning are language-specific and tightly constrain set of possible languages (Chomsky)
- Weak(er): There are language-specific biases toward learning possible languages (Mark Johnson)

Generalism (Empiricism)

- Language learned by general cognitive mechanisms shared with tasks like vision/motor control (Elman)
- Set of possible languages constrained by what the mechanisms can learn (Chater)

Controversies

Abstract (symbolic) representations

Infant grammar relatively abstract

- Rules, phrase structures

(Pinker)

Concrete representations

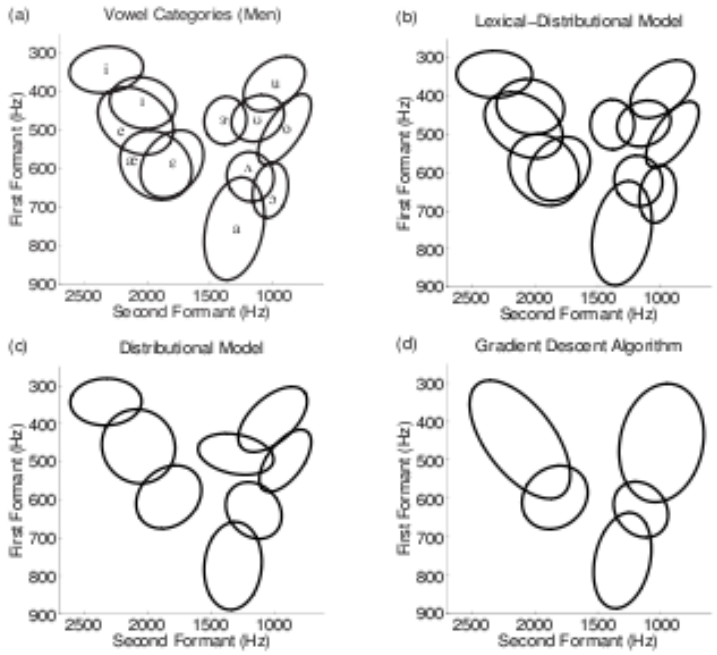
Infant grammar based on memorizing specific examples

- Constructions, exemplars, distributed representations (like neural nets)

(Tomasello)

- Difference here between theories that posit *adult* grammars based on exemplars...
- And those where infants “switch” in mid-development (Gleitman– the *tadpole-frog* problem)

Phonetics



(from Feldman et al '09)

How do infants learn to categorize speech sounds based on acoustics?

- Early models based on clustering and simple acoustics (McMurray et al)
- Influence from lexicon? (Feldman et al)
- Coarticulation a major issue

The lexicon

juwanttusiðəbʊk

ju • want • tu • si • ðə • bʊk

(Bernstein-Ratner corpus processed by Brent 99)

What evidence is used to segment words from fluent speech?

- Transition-based approaches (like Saffran) (Christiansen et al)
- Lexicalist approaches (distributional cues, syllable structure...) (Goldwater et al)
- Semantics? (Jones et al)
- Interaction with phonetics

Morphophonology

`bear` $\rightarrow$ `bear/z/, mat` $\rightarrow$ `mat/s/`

`wug` $\rightarrow$ `?`

Really several related problems, not going to cover all of them

- Phones to phonemes (Peperkamp et al)
- Morphological segmentation
- Word forms to paradigms
- Which features matter? What rules/constraints use them? (Wilson et al)
- Are there *underlying forms*? How are they represented?

Syntax

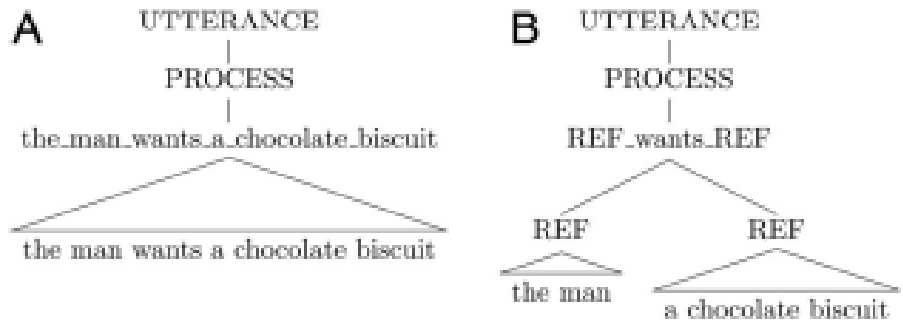


Fig. 1. Example analyses for the utterance *the man wants a chocolate biscuit* (*A*) Fully concrete. (*B*) Schema based.

(from Bannard et al)

What does early syntax look like? What biases are needed to learn it?

- Two cog-sci papers
  - (Perfors et al): against hardwired constraint that language must be hierarchical
  - (Bannard et al): adult-like grammar develops slowly over time
- One comp-sci paper: (Klein and Manning) beginning of modern NLP "grammar induction"
  - Controversial but very influential

Course objectives

If you take this course, I hope you will:

- Understand major trends in the literature at a high level
- Learn to read a statistical modeling paper, figure out what is going on and critique/evaluate the results
- Do a small project end-to-end
  - Come up with an interesting idea, do a directed literature search, implement a prototype, analyze the results

Course outline

- Two more lectures on statistical methods
  - Intro to the techniques used in the papers
- Six weeks of papers on modeling
  - A quick tour of the field
- Project proposals
- Four or five weeks of papers related to projects
  - Going into detail on things that interest you
- Project presentations

Paper discussions

Most classes will be discussions of papers:

- Everyone will read the paper
- Then post a question or comment on web discussion board the day before
- Discussion leader responsible for aggregating the questions...
- Moderating discussion of major issues raised by the class

Leader doesn't need to have all the answers...
But should make an effort to understand (may need to skim some cited work etc)
Should know *specifically* what they don't understand

Projects

Short written project proposal midway through the semester
Presentation at end

- Not expected to be a full conference paper
  - Although good projects might be extended into publications
- Should be conference-*style*
  - Including motivation, references, model, experiment and analysis
- Must have an implementational component
  - Could be very ambitious...
  - Or fairly simple
  - Can use other peoples' software
- Doesn't have to "work" (get an interesting result on a real dataset)
  - In this case, analysis should be specific about why not

What you need

- Basic linguistics... intro phonetics/phonology/syntax
  - Dev. linguistics not my field; I'd be glad if you know it
  - But course should be accessible without
- Basic probability (independence, marginalization, Bayes' rule)
  - If that's all you have, course probably survivable but *difficult*... but you'll learn a lot!
  - Will be easier if you know: graphical models, EM algorithm, conjugate priors, Gibbs sampling (next 2 lectures)
- Project should involve getting your hands dirty...
  - If you can't program *at all*, should audit
  - My language for examples: python. Also fine: R, Matlab, perl
    - C, C++, Java fine but often overkill!
- Visitors welcome...
  - Don't have to take the course or even come all the time