Word Order Acquisition from Immature Representations: A Model for Noise-Tolerant Learning of Deterministic Grammars

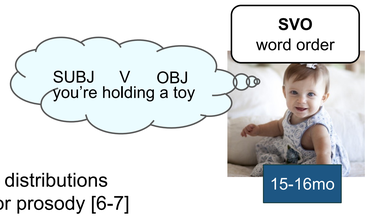
Laurel Perkins and Tim Hunter • University of California, Los Angeles

Learning from Noisy Data

Learners use developing grammatical knowledge to parse & learn from their data
► How do they generalize accurately from **immature representations** of input?

Case Study: Word Order

Knowledge of canonical word order is acquired in infancy [1-4]

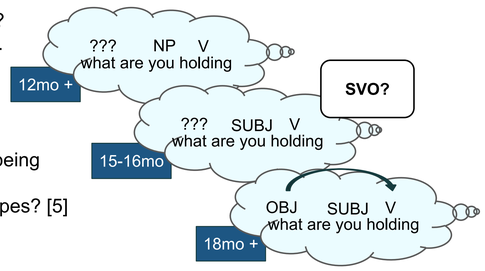


Possible mechanisms:

- Learning from noun phrase & verb distributions
- Bootstrapping from semantics [5] or prosody [6-7]

Problem: reliably identifying moved clause arguments takes time [4, 8-11]

(1) What are you holding?
(2) That's the dog we like.
(3) You're being hugged (by your sister).



► How do children avoid being misled by "noise" from non-canonical clause types? [5]

Proposal: Input Filtering

Expect that data are a noisy realization of a deterministic underlying system, and learn to filter noise
► Previously applied to learning of verb transitivity classes [12]
► **Current work: this mechanism generalizes** to more complex rule systems

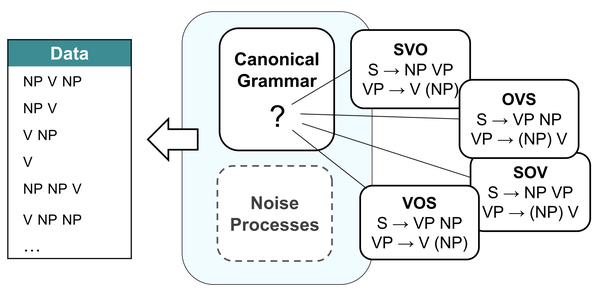
Our Model

Observes strings of imperfectly-identified NPs and Vs, considers 4-way choice of canonical word order

- Grammar deterministically places subject before/after VP, object before/after V
- Some parts of strings are generated by "noise" processes: unknown grammatical phenomena that appear to insert, delete, or swap arguments

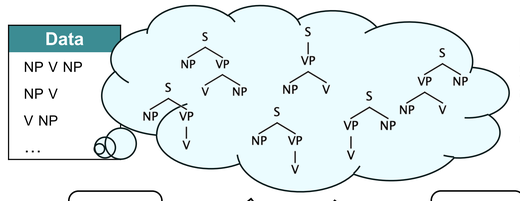
Bayesian joint inference to select a canonical grammar + filter parameters

- What do the data from the canonical grammar look like?
- What do the data from noise look like?
- What is the right division into signal vs. noise?

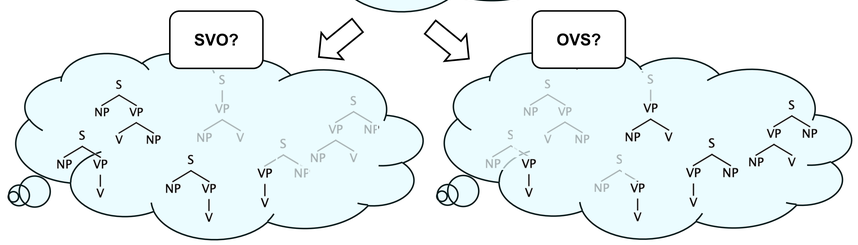


What does Filtering Look Like?

From strings of NPs and Vs, make a noisy guess about underlying tree structure



Decide which parts of guessed trees are **signal,** and which are **noise,** for an underlying word order grammar



► It is possible to **learn** how to divide up the data into signal vs. noise, without knowing ahead of time how much noise there is, or what its properties are

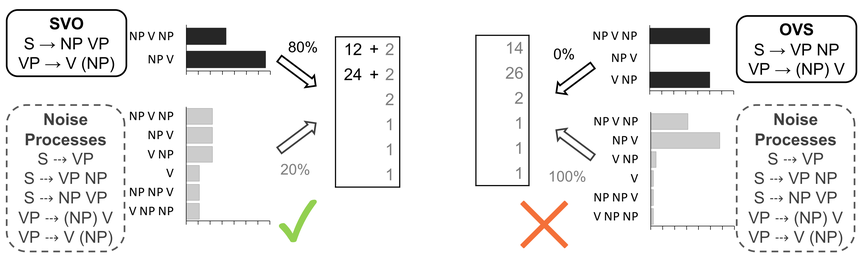
Toy Example

How might these data have arisen partially from a word order grammar distribution, and partially from the noise distribution?

Data: 45 strings	
NP V NP	14
NP V	26
V NP	2
V	1
NP NP V	1
V NP NP	1

SVO? OVS? SOV? VOS?

Two solutions (of many):



- Costly to analyze too much of the data as noise: too many degrees of freedom
- Simpler solution: attribute skewed data to restrictive word order grammar whenever possible

Simulations: Child-Directed English and French

50-sentence datasets sampled from Eve & Lyon CHILDES corpora [13-14]

- Strings of Vs and NPs imperfectly identified from functional cues [15-17]
- Cues for NPs: is a full pronoun, or follows a determiner
- Cues for Vs: follows an auxiliary

From these strings, our model infers posterior probability distribution over underlying trees and word order grammars

English		French	
0.36	NP V	0.48	NP V
0.20	V	0.21	NP V NP
0.20	NP V NP	0.13	V
0.17	V NP	0.05	NP NP V
0.04	NP V NP NP	0.03	NP V NP NP
0.03	V NP NP	0.03	V NP

Fig. 1 Most frequent string types

Results

Our Model

Learner successfully assigns SVO highest posterior probability in both languages

- Even though data cannot be produced by any single word order grammar, without noise

► Filter works, and **filter can be learned** from distributions in the data

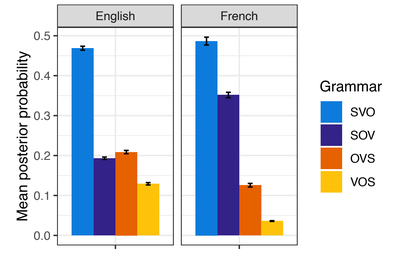


Fig. 2 Posterior distribution over word order grammars

Comparison: "Fully-Flexible" Learner

No 4-way choice of a canonical word order grammar: one grammar, all rules possible with some probability [18]

- Collapses distinction in our model between rules for canonical and non-canonical structures
- Learning canonical word order means identifying that some rules have probabilities near zero

Two variants: with and without an explicit bias to regularize (push probabilities towards zero/one) [18-21]

- Learner without bias to regularize infers distributions that mirror its noisy data
- Learner with bias to regularize gives high probability to several canonical word orders

► Useful to have a hypothesis space comprising restrictive options

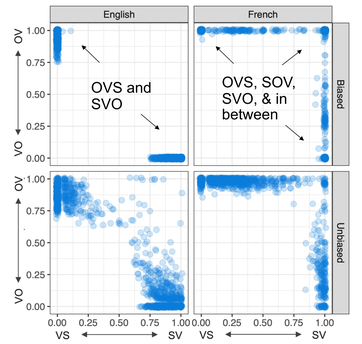


Fig. 3 Posterior distribution over subject and object positions in trees, fully-flexible learner

Discussion

We find that input filtering can in principle enable acquisition of basic word order from noisy data

- From imperfectly-identified NP and V distributions alone, our model learns to separate evidence for canonical word order from the distorting effects of "noise" processes
- It does so without knowing ahead of time what noise looks like, or how much there is

Restrictive options in the learner's hypothesis space allow successful filtering

- Each word order grammar allows only a certain combination of rules
- Preference emerges to use these when possible, rather than analyzing everything as noise

► Provides a novel mechanism for **regularization** in grammar learning [18-21]

► Grammar leads you to expect regularities in your data
► Filtering allows you to find them

Acknowledgments

Our thanks to Xinyue Cui, Naomi Feldman, Jeffrey Lidz, Shalinee Maitra, the UCLA Psycholinguistics/Computational Linguistics Seminar, and the University of Maryland Linguistics General Meeting for helpful feedback and assistance.

Contact: perkinsl@ucla.edu, timhunter@ucla.edu. **Selected references:** [1] Hirsh-Pasek & Golinkoff 1996. The intermodal preferential looking paradigm. *Language, Speech, and Communication*. [2] Gertner et al. 2006. Learning words and rules. *Psychol Sci.* [3] Lidz, White & Baier 2017. The role of incremental parsing in syntactically conditioned word learning. *Cogn Psychol*. [4] Perkins & Lidz 2020. Filler-gap dependency comprehension at 15 months. *Lang Acq*. [5] Pinker 1984. *Language Learnability and Language Development*. [6] Morgan 1986. *From Simple Input to Complex Grammar*. [7] Christophe et al. 2008. Bootstrapping lexical and syntactic acquisition. *Lang and Speech*. [8] Perkins & Lidz 2021.18-month-old infants represent non-local syntactic dependencies. *PNAS*. [9] Gagliardi, Mease & Lidz 2016. Discontinuous development in the acquisition of filler-gap dependencies. *Lang Acq*. [10] Seidl, Hollich, & Jusczyk 2003. Early understanding of subject and object wh-questions. *Infancy*. [11] Dautriche et al. 2014. Toddlers default to canonical surface-to-meaning mapping when learning verbs. *Child Dev*. [12] Perkins et al. 2022. The power of ignoring. *Cogn Sci*. [13] Brown 1973. *A first language: The early stages*. [14] Demuth & Tremblay 2008. Prosodically-conditioned variability in children's production of French determiners. *J. Child Lang*. [15] Babineau & Christophe 2022. Preverbal infants' sensitivity to grammatical dependencies. *Infancy*. [16] Shi & Melançon 2010. Syntactic categorization in French-learning infants. *Infancy*. [17] Hicks, Maye, & Lidz 2007. The role of function words in infants' syntactic categorization of novel words. *LSA*. [18] Johnson et al. 2007. Bayesian inference for PCFGs via Markov chain Monte Carlo. *ACL*. [19] Reali & Griffiths 2009. The evolution of frequency distributions. *Cognition*. [20] Culbertson & Smolensky 2012. A Bayesian model of biases in artificial language learning. *Cogn Sci*. [21] Hudson Kam & Newport 2009. Getting it right by getting it wrong. *Cogn Psychol*.