Noise-tolerant learning as selection among deterministic grammatical hypotheses

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

Infants acquire their language's basic word order from input containing a mixture of canonical and non-canonical sentence types [1-4]

(1) You're holding a toy.
(2) What are you holding?
(3) That's the dog we like.
(4) You're being hugged.

SVO?

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

Distributions in child-directed speech are potentially misleading

► How do children avoid being misled by "noise" from non-canonical clause types? [4-8]

Fig. 1 Most frequent string types in Eve and Lyon CHILDES corpora [9-10]; NPs and Vs imperfectly identified from functional cues [11-13]

Proposal: Input Filtering

Two Possible Solutions

Regularization: explicit numerical bias against encoding full variability of data: prefer hypotheses that are heavily skewed [14-16]

Filtering: expect that data are a noisy realization of a deterministic underlying system, and learn to separate signal from noise [8]

► **Finding: filtering works better** in this learning domain

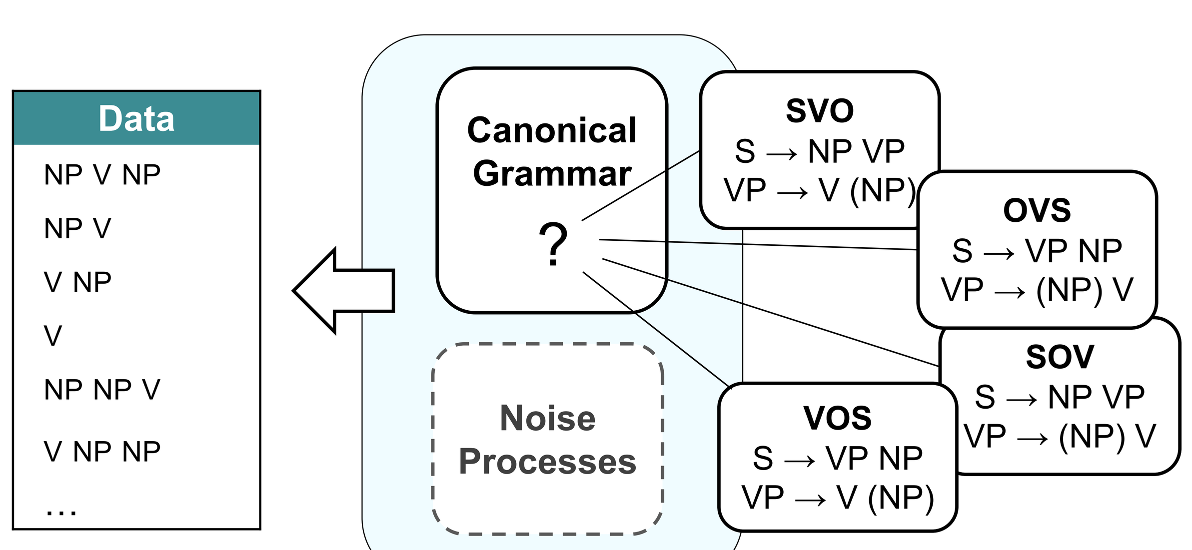
Our Model

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

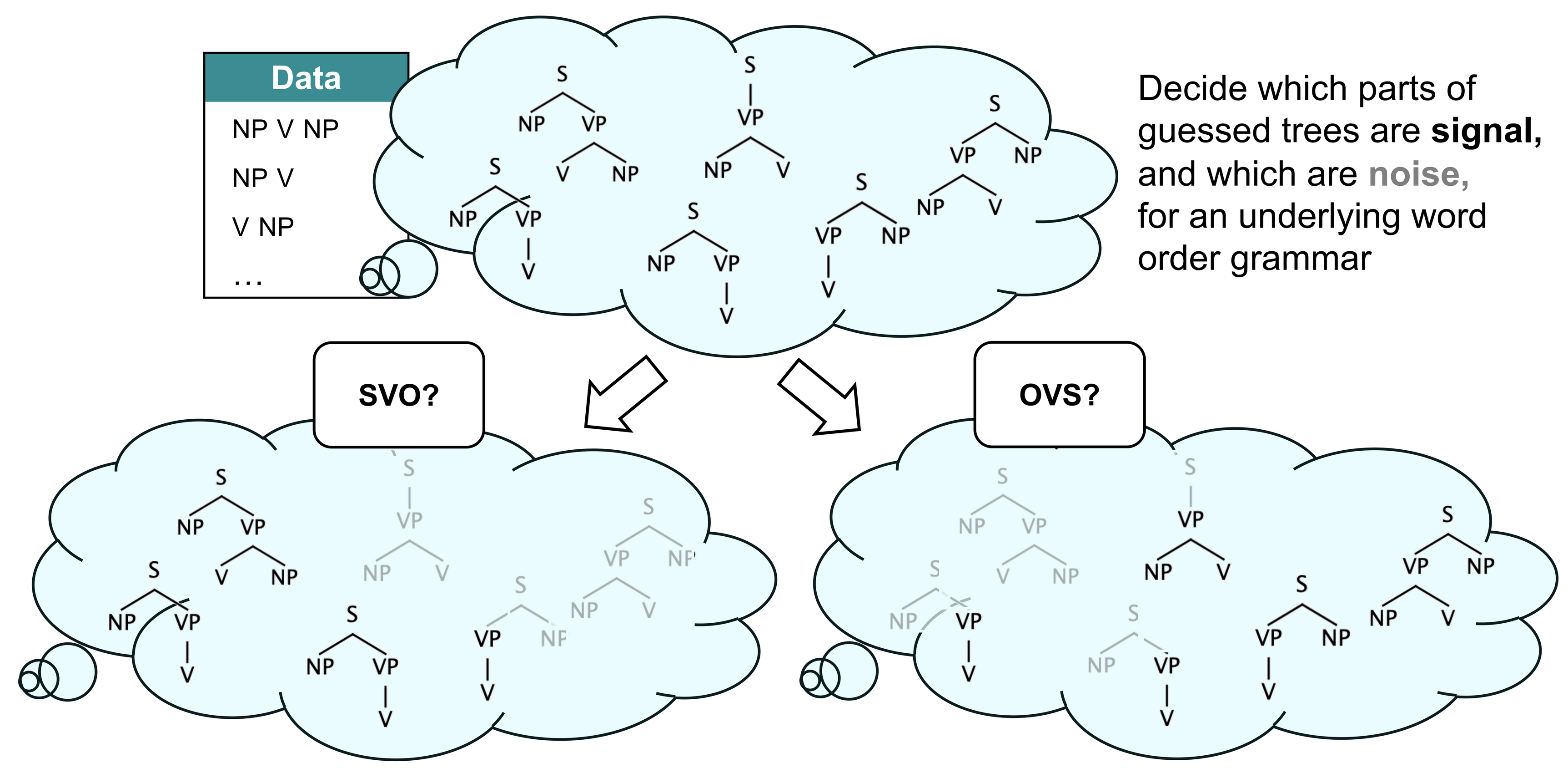
Bayesian joint inference to select a canonical grammar + filter parameters

- 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

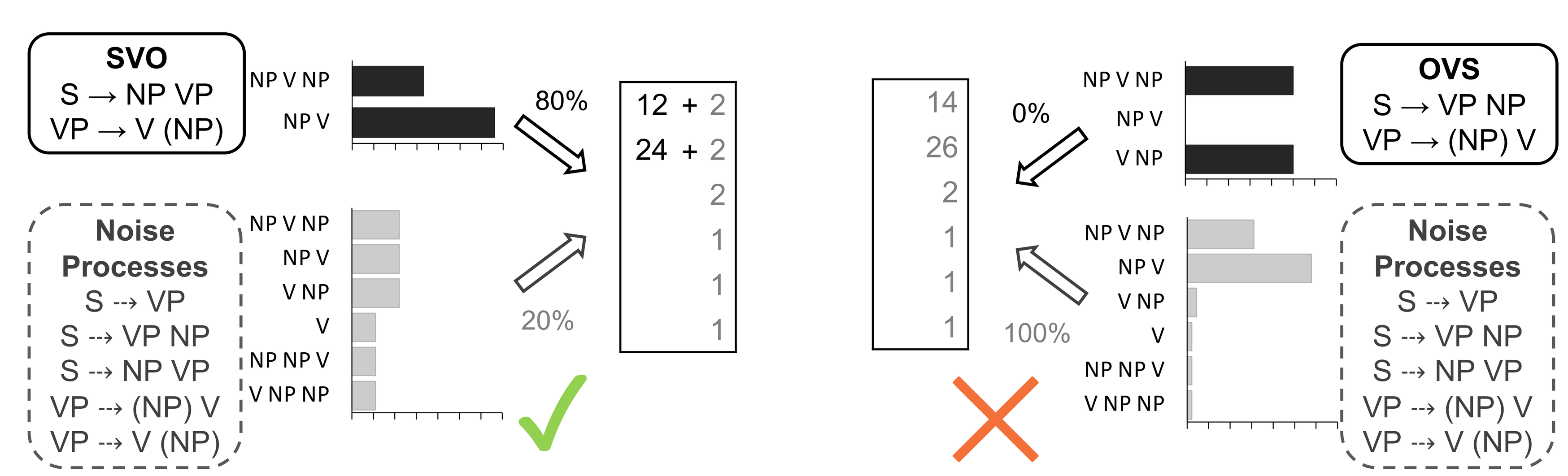
Toy Example

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

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

Results: Child-Directed Speech

Simulations on 50-sentence datasets of NP-V strings, sampled from corpora of child-directed English and French (Fig. 1)

- 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

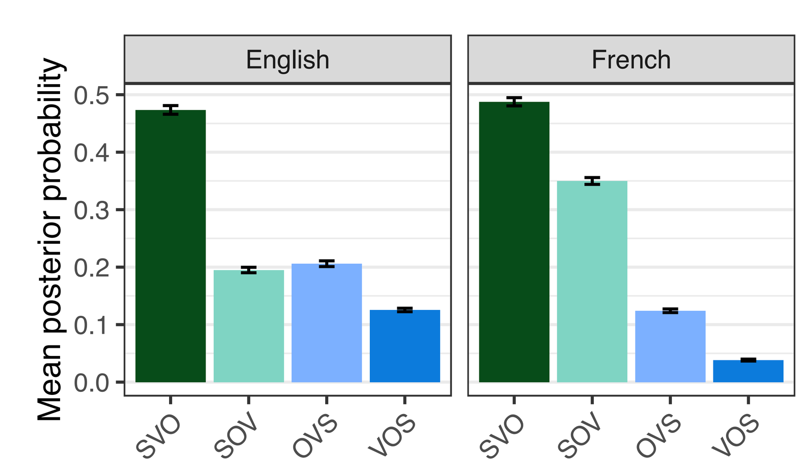


Fig. 2 Posterior distribution over word order grammars

Model Comparisons

"Fully-Flexible" Learner

No 4-way choice of a canonical word order 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) [15-17]

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

► Useful to have a hypothesis space with restrictive grammatical options

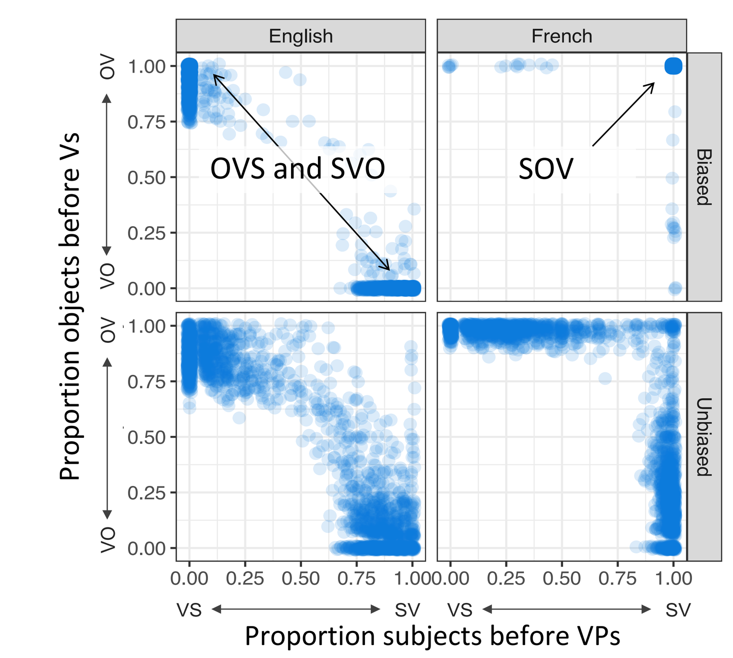


Fig. 3 Posterior distribution over subject and object positions in trees, fully-flexible learner

A Data-Coverage Heuristic

Simpler alternative: select the grammar that covers the most data

- E.g., core rules of SVO grammar generate 56% of English data, more than any other grammar

Comparison: version of our learner with an 8-way choice among grammars

- 4 options from original model that fix both subject and object position
- 4 less restrictive options that only fix one argument position, and allow the other to vary

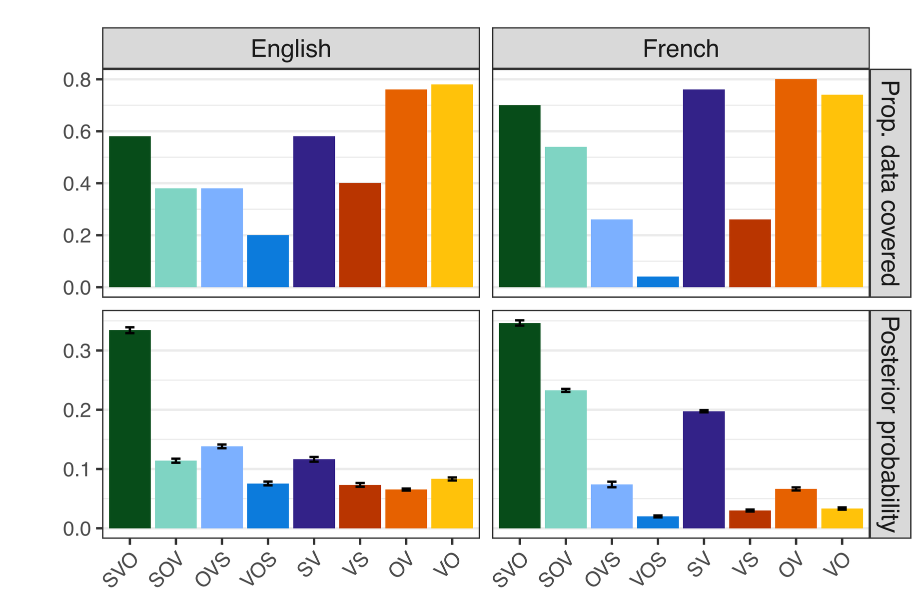


Fig. 4 Eight-way hypothesis space: proportion data coverage vs. model's posterior distribution

Our learner still successfully assigns SVO highest posterior probability in both languages, even though more flexible grammars cover more of the data

► Emergent preference for most restrictive hypothesis that fits the data

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 alternative to **regularization** in grammar learning

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

Acknowledgments: Thanks to Xinyue Cui, Naomi Feldman, Jeffrey Lidz, Shalinee Maitra, and audiences at the UCLA Psycholinguistics/Computational Linguistics Seminar and BUCLD 2022.
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] Perkins & Lidz 2021.18-month-old infants represent non-local syntactic dependencies. *PNAS*. [6] Gagliardi, Mease & Lidz 2016. Discontinuous development in the acquisition of filler-gap dependencies. *Lang Acq.* [7] Dautriche et al. 2014. Toddlers default to canonical surface-to-meaning mapping when learning verbs. *Child Dev*. [8] Perkins et al. 2022. The power of ignoring. *Cogn Sci.* [9] Brown 1973. *A first language: The early stages.* [10] Demuth & Tremblay 2008. Prosodically-conditioned variability in children's production of French determiners. *J. Child Lang.* [11] Babineau & Christophe 2022. Preverbal infants' sensitivity to grammatical dependencies. *Infancy.* [12] Shi & Melançon 2010. Syntactic categorization in French-learning infants. *Infancy.* [13] Hicks, Maye, & Lidz 2007. The role of function words in infants' syntactic categorization of novel words. *LSA.* [14] Reali & Griffiths 2009. The evolution of frequency distributions. *Cognition.* [15] Culbertson & Smolensky 2012. A Bayesian model of biases in artificial language learning. *Cogn Sci.* [16] Hudson Kam & Newport 2009. Getting it right by getting it wrong. *Cogn Psychol.* [17] Johnson et al. 2007. Bayesian inference for PCFGs via Markov chain Monte Carlo. *ACL.*
Contact: perkinsl@ucla.edu, timhunter@ucla.edu