Learning Recursion

Bob Frank
Department of Linguistics
Yale University

Recursion and human language

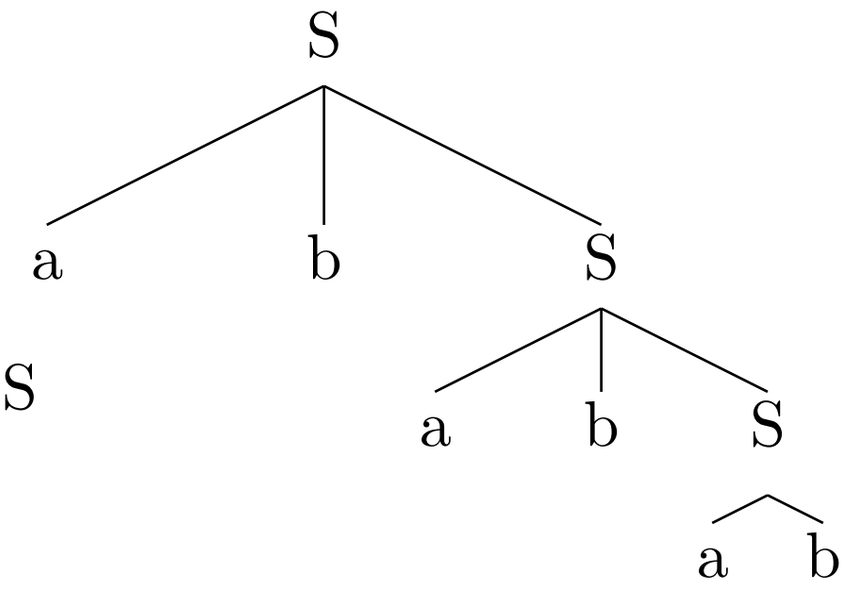
Hauser, Chomsky and Fitch (2002)

“We hypothesize that FLN only includes recursion and is the only uniquely human component of the faculty of language.”

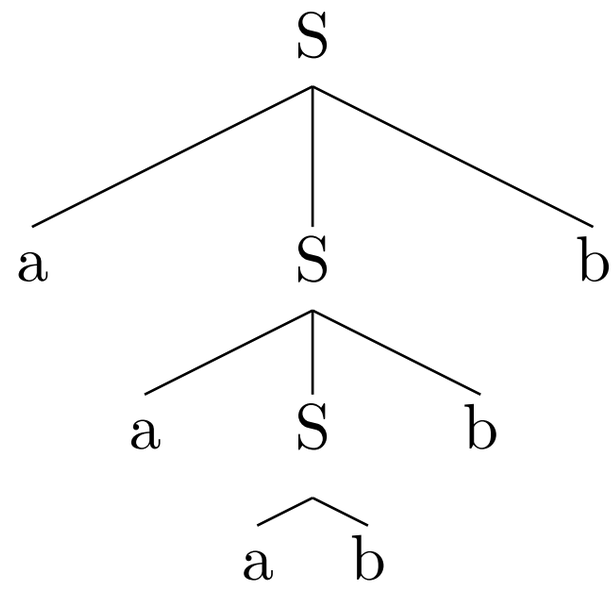
Recursion and Human Language

- What counts as recursion?

$$S \rightarrow a\ b\ S$$
$$S \rightarrow a\ b$$



$$S \rightarrow a\ S\ b$$
$$S \rightarrow a\ b$$



Uniquely human?

Computational Constraints on Syntactic Processing in a Nonhuman Primate

W. Tecumseh Fitch[1]* and Marc D. Hauser[2]

The capacity to generate a limitless range of meaningful expressions from a finite set of elements differentiates human language from other animal communication systems. Rule systems capable of generating an infinite set of outputs ("grammars") vary in generative power. The weakest possess only local organizational principles, with regularities limited to neighboring units. We used a familiarization/discrimination paradigm to demonstrate that monkeys can spontaneously master such grammars. However, human language entails more sophisticated grammars, incorporating hierarchical structure. Monkeys tested with the same methods, syllables, and sequence lengths were unable to master a grammar at this higher, "phrase structure grammar" level.

Uniquely human?

- Familiarization for 20 minutes to 60 strings of one of two forms (where n=2 or 3)

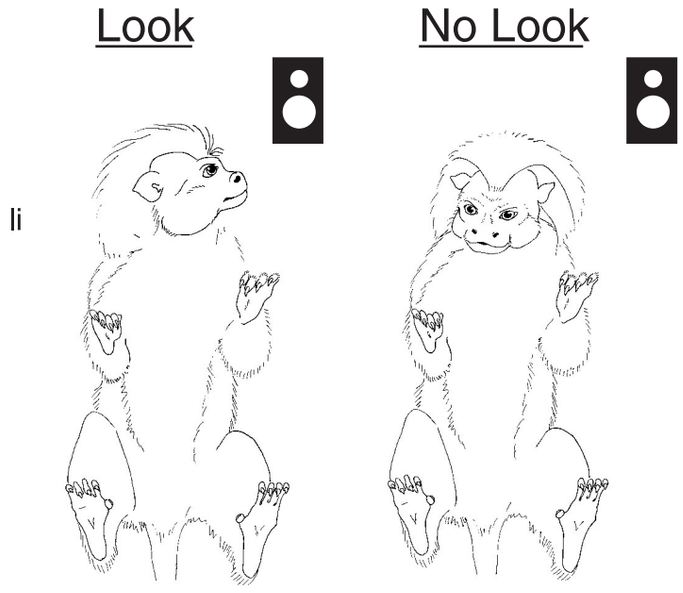
Finite State Grammar $(AB)^n$

AB AB **no** li **ba** pa
AB AB AB **la** pa **wu** mo **no** li

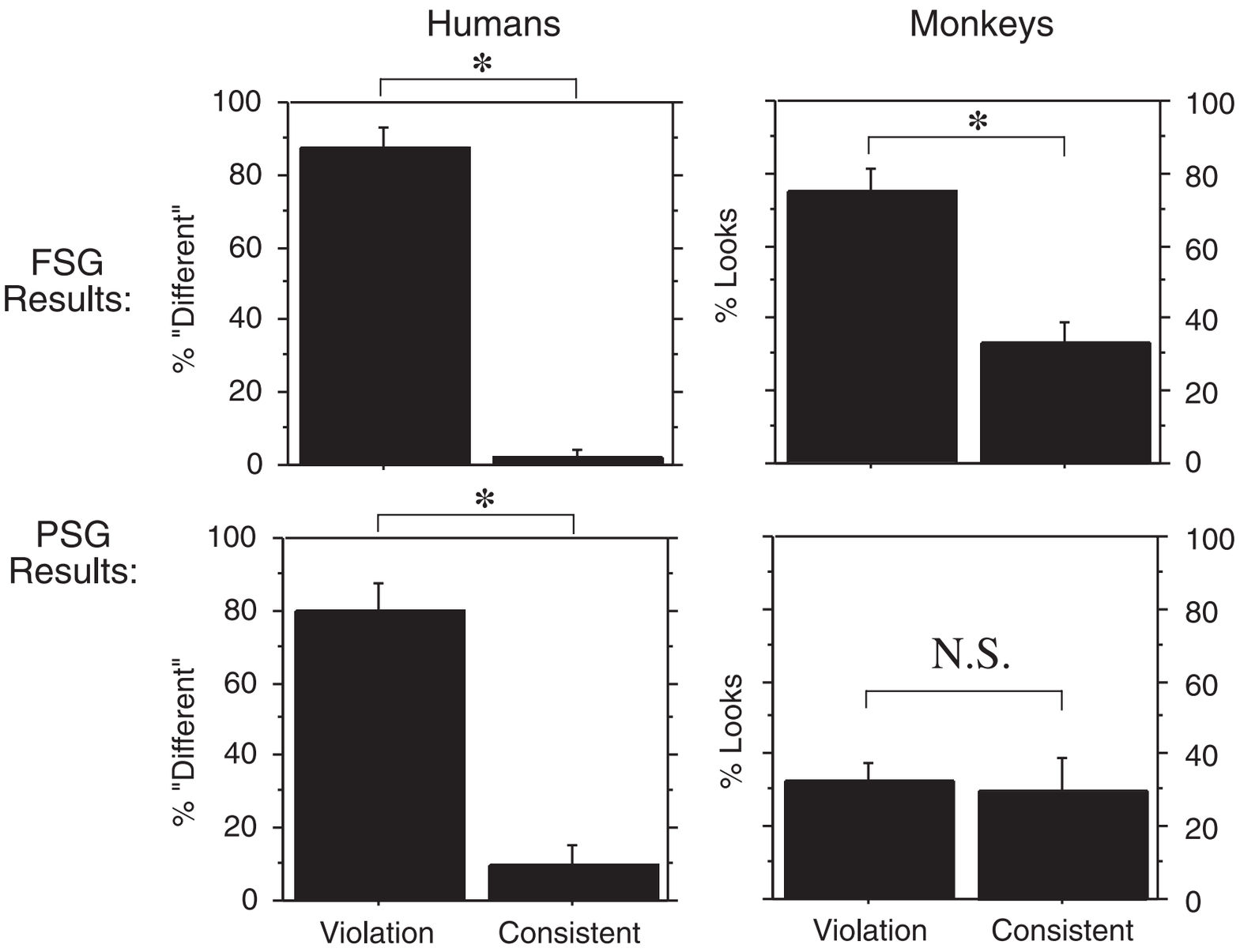
Phrase Structure Grammar: A^nB^n

AA BB **yo la** pa do
AAA BBB **ba la tu** li pa ka

- Test for generalization to novel strings



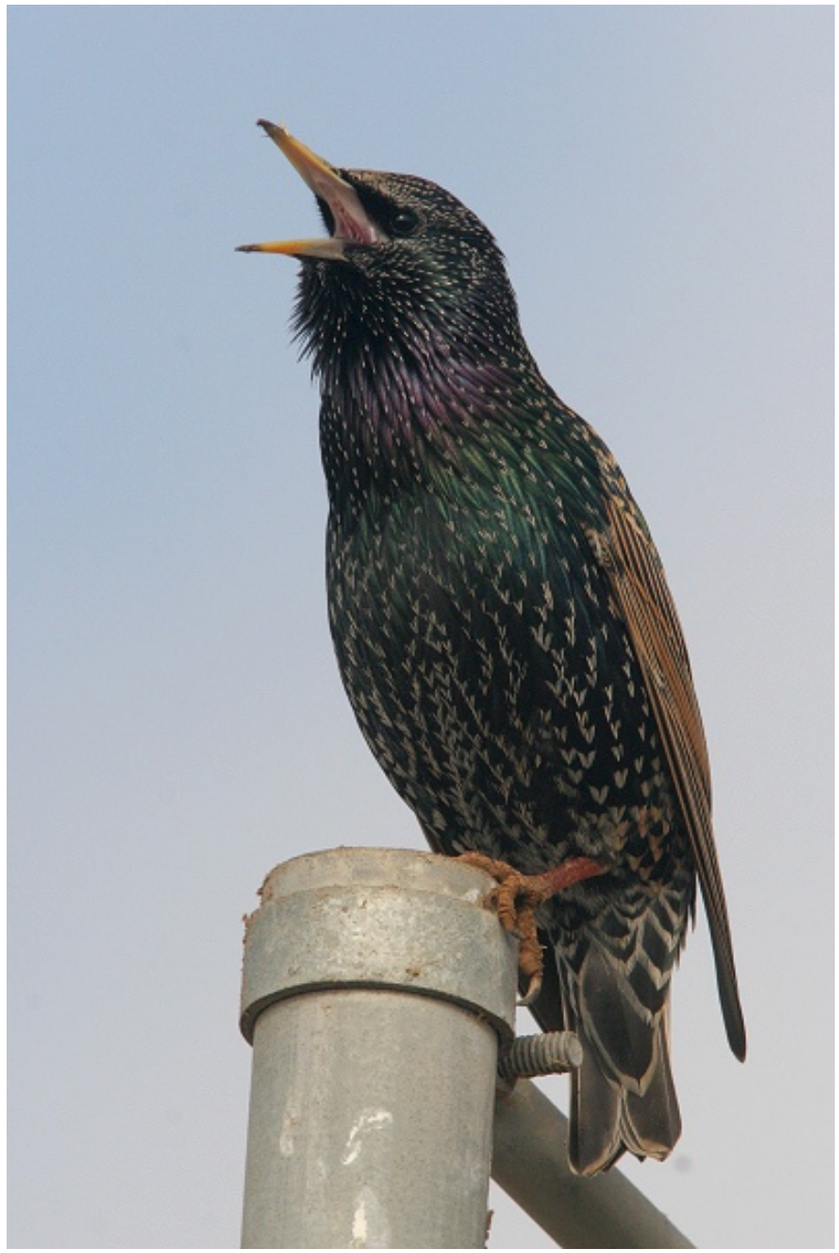
Tamarin results



Uniquely human?

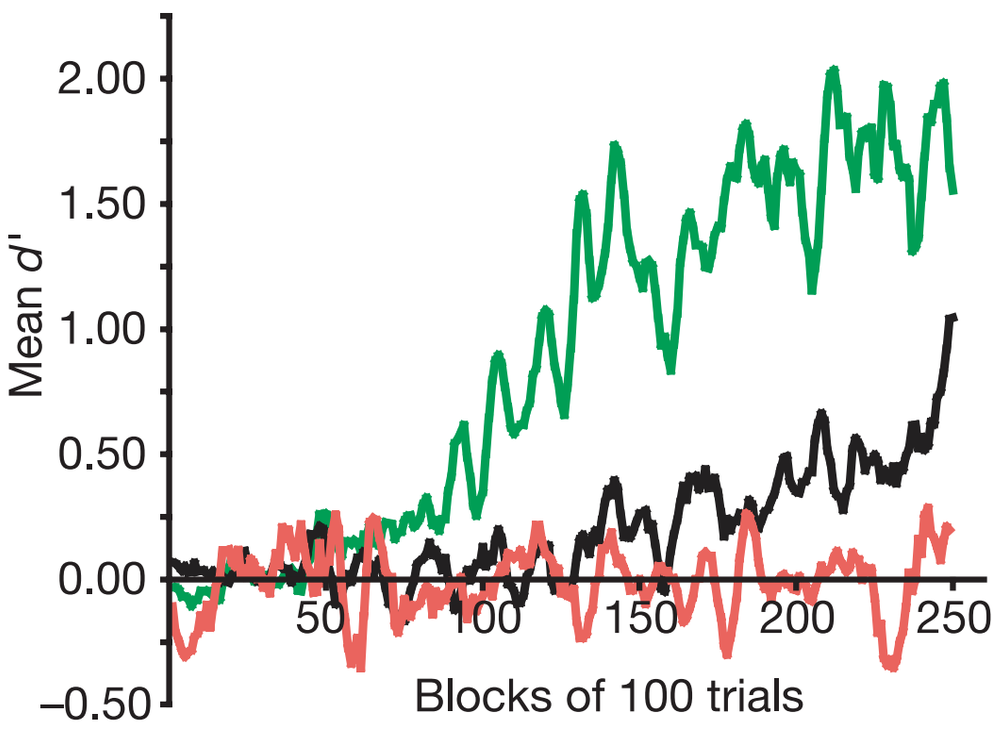
Recursive syntactic pattern learning by songbirds

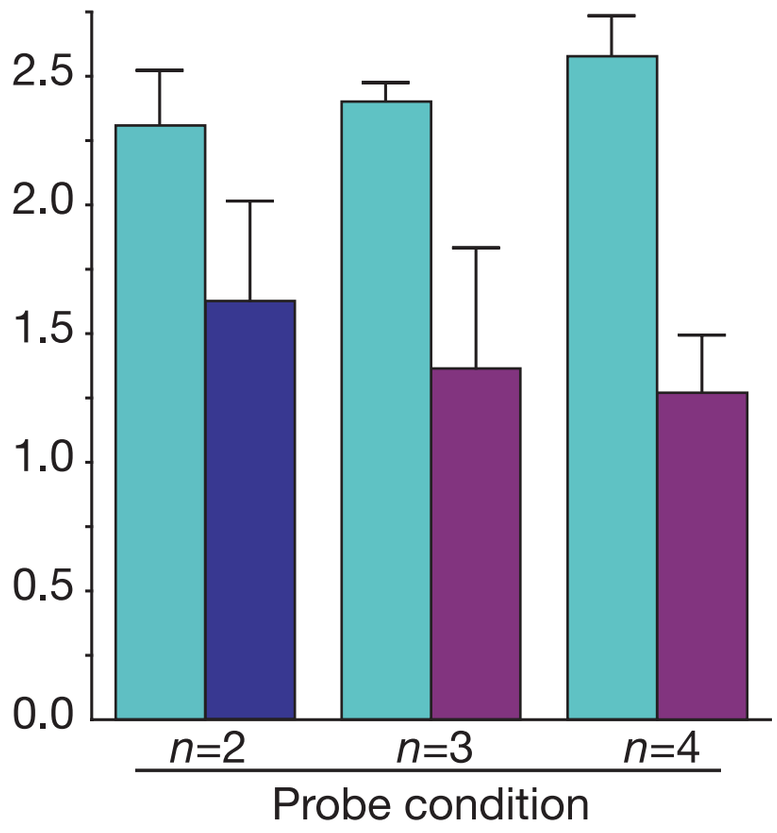
Timothy Q. Gentner[1]†, Kimberly M. Fenn[2], Daniel Margoliash[1,2] & Howard C. Nusbaum[2]



Humans regularly produce new utterances that are understood by other members of the same language community[1]. Linguistic theories account for this ability through the use of syntactic rules (or generative grammars) that describe the acceptable structure of utterances[2]. The recursive, hierarchical embedding of language units (for example, words or phrases within shorter sentences) that is part of the ability to construct new utterances minimally requires a 'context-free' grammar[2,3] that is more complex than the 'finite-state' grammars thought sufficient to specify the structure of all non-human communication signals. Recent hypotheses make the central claim that the capacity for syntactic recursion forms the computational core of a uniquely human language faculty[4,5]. Here we show that European starlings (*Sturnus vulgaris*) accurately recognize acoustic patterns defined by a recursive, self-embedding, context-free grammar. They are also able to classify new patterns defined by the grammar and reliably exclude agrammatical patterns. Thus, the capacity to classify sequences from recursive, centre-embedded grammars is not uniquely human. This finding opens a new range of complex syntactic processing mechanisms to physiological investigation.

Starlings and Recursion





Do Humans Really Learn $A^n B^n$ Artificial Grammars From Exemplars?

Jean-Rémy Hochmann, Mahan Azadpour, Jacques Mehler

Department of Cognitive Neuroscience, International School of Advanced Studies

Received 5 April 2007; received in revised form 10 December 2007; accepted 11 December 2007

Abstract

An important topic in the evolution of language is the kinds of grammars that can be computed by humans and other animals. Fitch and Hauser (F&H; 2004) approached this question by assessing the ability of different species to learn 2 grammars, $(AB)^n$ and $A^n B^n$. $A^n B^n$ was taken to indicate a *phrase structure grammar,* eliciting a center-embedded pattern. $(AB)^n$ indicates a grammar whose strings entail only local relations between the categories of constituents. F&H's data suggest that humans, but not tamarin monkeys, learn an $A^n B^n$ grammar, whereas both learn a simpler $(AB)^n$ grammar (Fitch & Hauser, 2004). In their experiments, the A constituents were syllables pronounced by a female voice, whereas the B constituents were syllables pronounced by a male voice. This study proposes that what characterizes the $A^n B^n$ exemplars is the distributional regularities of the syllables pronounced by either a male or a female rather than the underlying, more abstract patterns. This article replicates F&H's data and reports new controls using either categories similar to those in F&H or less salient ones. This article shows that distributional regularities explain the data better than grammar learning. Indeed, when familiarized with $A^n B^n$ exemplars, participants failed to discriminate $A^3 B^2$ and $A^2 B^3$ from $A^n B^n$ items, missing the crucial feature that the number of As must equal the number of Bs. Therefore, contrary to F&H, this study concludes that no syntactic rules implementing embedded nonadjacent dependencies were learned in these experiments. The difference between human linguistic abilities and the putative precursors in monkeys deserves further exploration.

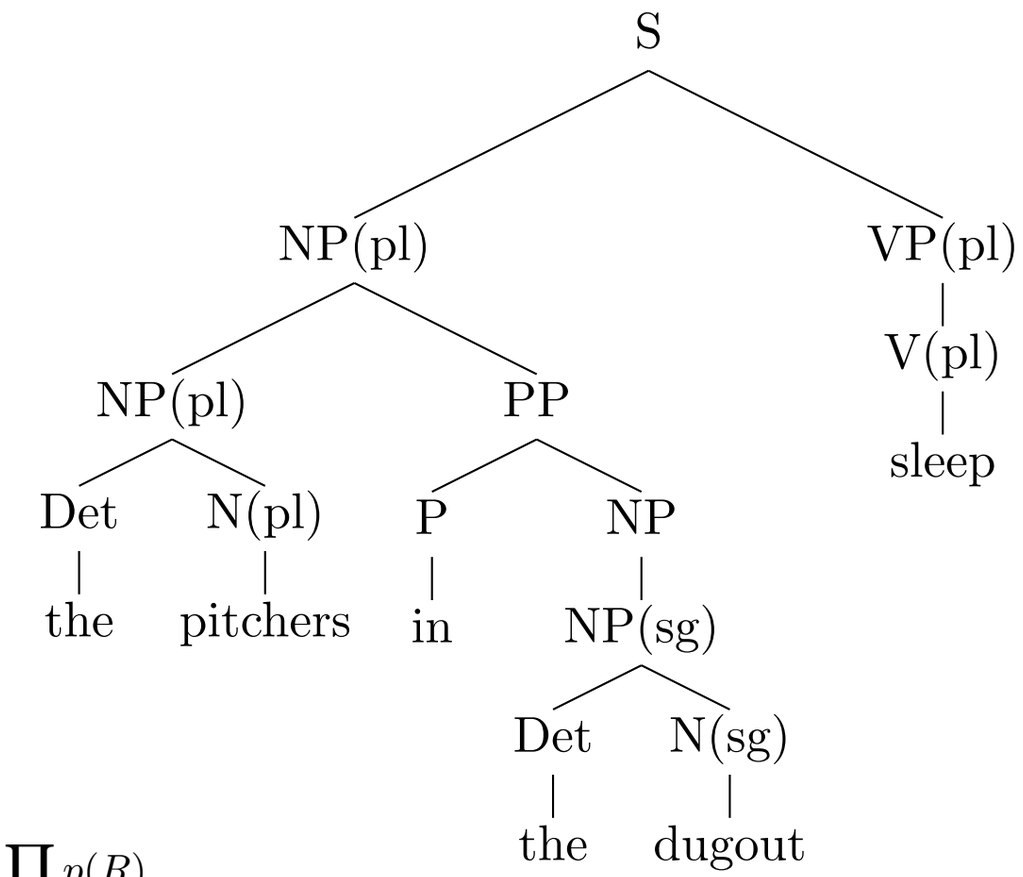
What are these experimental results telling us?

- Are starlings capable of recursion, while cotton-top tamarins not?
- Are humans capable of recursion?
- Is a difference in the experimental methodology responsible for the outcome?
- Would a rational learner in these experiments to identify recursion?

Statistical Models of Grammar

- Probabilistic Context-Free Grammars (PCFGs)

$$
\begin{array}{rcl}
\text{S} & \rightarrow & \text{NP(sg) VP(sg)} \\
\text{S} & \rightarrow & \text{NP(pl) VP(pl)} \\
\text{NP(sg)} & \rightarrow & \text{Name} \\
\text{NP(sg)} & \rightarrow & \text{Det N(sg)} \\
\text{NP(sg)} & \rightarrow & \text{NP(sg) PP} \\
\text{NP(pl)} & \rightarrow & \text{Det N(sg)} \\
\text{NP(pl)} & \rightarrow & \text{NP(pl) PP} \\
\text{NP} & \rightarrow & \text{NP(sg)} \\
\text{NP} & \rightarrow & \text{NP(pl)} \\
\text{PP} & \rightarrow & \text{P NP} \\
\text{VP(pl)} & \rightarrow & \text{V(pl)} \\
\text{VP(pl)} & \rightarrow & \text{V(pl) NP} \\
\text{VP(pl)} & \rightarrow & \text{V(pl) S}
\end{array}
$$



$$
\begin{aligned}
p(T) &= \prod_R p(R) \\
&= p(\text{S} \rightarrow \text{NP VP}) \times p(\text{NP(pl)} \rightarrow \text{NP(pl) PP}) \times p(\text{NP(pl)} \rightarrow \text{Det N}) \times \ldots \\
&= .5 \times .1 \times .9 \times 1 \times .5 \times .4 \times .3
\end{aligned}
$$

$$
p(w_1 w_2 \ldots w_n) = \sum_T p(T | y(T) = w_1 w_2 \ldots w_n)
$$

Statistical properties of recursion

- A PCFG of the following form will generate the relevant type of recursive strings:

$$\begin{aligned} (\mathrm{p})\ \mathrm{S} &\rightarrow \mathrm{a}\ \mathrm{S}\ \mathrm{b} \\ (\mathrm{1\text{-}p})\ \mathrm{S} &\rightarrow \mathrm{a}\ \mathrm{b} \end{aligned}$$

- The distribution of strings generated by such a grammar has a distinctive property: shorter strings are always more frequent than longer strings.

$$P(\mathrm{a}^k\mathrm{b}^k) = p^{k-1}(1-p)$$

- In contrast, there is no recursive PCFG that can generate a set of strings in which each length is equally likely. Instead, such a set could only be generated by a non-recursive grammar:

$$\begin{aligned} (\mathrm{p})\ \mathrm{S} &\rightarrow \mathrm{a}\ \mathrm{b} \\ (\mathrm{q})\ \mathrm{S} &\rightarrow \mathrm{a}\ \mathrm{a}\ \mathrm{b}\ \mathrm{b} \\ (\mathrm{r})\ \mathrm{S} &\rightarrow \mathrm{a}\ \mathrm{a}\ \mathrm{a}\ \mathrm{b}\ \mathrm{b}\ \mathrm{b} \\ (\mathrm{1\text{-}(p+q+r)})\ \mathrm{S} &\rightarrow \mathrm{a}\ \mathrm{a}\ \mathrm{a}\ \mathrm{a}\ \mathrm{b}\ \mathrm{b}\ \mathrm{b}\ \mathrm{b} \end{aligned}$$

Statistical properties of recursion

- We find this patterning in natural language
  - Sentential embedding in the Brown corpus:

	no embedding	1 level of embedding	2 levels of embedding	3 levels of embedding or more
# of sentences	15366	5874	826	113
percentage	0.69	0.26	0.04	<0.01

$$(.69)\ \mathrm{S} \rightarrow \ldots\ \mathrm{V}\ \ldots$$
$$(.31)\ \mathrm{S} \rightarrow \ldots\ \mathrm{V}\ \mathrm{S}$$

overgenerates longer examples

Statistical grammar learning

- Work in computational linguistics has addressed the problem of grammar learning from a corpus of utterances.
  - Given a corpus of data D, and a set of PCFG grammar rules
  - We might want to find the set of rule probabilities that maximizes the likelihood of the data (*maximum likelihood learning*)

$$\text{argmax}_\theta \; p(D|\theta) = \text{argmax}_\theta \sum_T p(D|T)p(T|\theta)$$

- This task is not trivial, because we don't know what the right trees are for the sentences. However, there are algorithms which allow us to solve this problem (inside-outside).
- Can such a learner detect the signature properties of recursion just discussed?

Statistical learning of recursion

- Inside-Outside algorithm trained on corpus of sentences of the form A^nB^n (10 different words could fill in A and B), generated either by

1. a recursive PCFG, with non-recursive rule probability 1/3; or

$$(.666)\ \mathrm{S} \rightarrow \mathrm{A\ S\ B}$$
$$(.333)\ \mathrm{S} \rightarrow \mathrm{A\ B}$$

2. a non-recursive PCFG with equal rule probabilities for each string length (up to n=3).

$$(.333)\ \mathrm{S} \rightarrow \mathrm{A\ B}$$
$$(.333)\ \mathrm{S} \rightarrow \mathrm{A\ A\ B\ B}$$
$$(.333)\ \mathrm{S} \rightarrow \mathrm{A\ A\ A\ B\ B\ B}$$

- The initial set of rules was as follows:

$$\mathrm{S} \rightarrow \mathrm{A\ S\ B}$$
$$\mathrm{S} \rightarrow \mathrm{A\ B}$$
$$\mathrm{S} \rightarrow \mathrm{A\ A\ B\ B}$$
$$\mathrm{S} \rightarrow \mathrm{A\ A\ A\ B\ B\ B}$$
$$\mathrm{S} \rightarrow \mathrm{A\ A\ A\ A\ B\ B\ B\ B}$$

Statistical learning of recursion

- Results for statistically recursive training data:

$$
\begin{aligned}
(.648)\ \mathrm{S} &\rightarrow \mathrm{A\ S\ B} \\
(.325)\ \mathrm{S} &\rightarrow \mathrm{A\ B} \\
(.005)\ \mathrm{S} &\rightarrow \mathrm{A\ A\ B\ B} \\
(.018)\ \mathrm{S} &\rightarrow \mathrm{A\ A\ A\ B\ B\ B} \\
(.004)\ \mathrm{S} &\rightarrow \mathrm{A\ A\ A\ A\ B\ B\ B\ B}
\end{aligned}
$$

- Results for statistically non-recursive training data:

$$
\begin{aligned}
(0)\ \mathrm{S} &\rightarrow \mathrm{A\ S\ B} \\
(.370)\ \mathrm{S} &\rightarrow \mathrm{A\ B} \\
(.300)\ \mathrm{S} &\rightarrow \mathrm{A\ A\ B\ B} \\
(.330)\ \mathrm{S} &\rightarrow \mathrm{A\ A\ A\ B\ B\ B} \\
(0)\ \mathrm{S} &\rightarrow \mathrm{A\ A\ A\ A\ B\ B\ B\ B}
\end{aligned}
$$

Statistical learning of recursion

- As far as I can tell from the descriptions in the papers, previous experiments used such a statistically non-recursive distribution. As a result, the cotton-top tamarins are behaving perfectly rationally, even if they have the possibility of representing grammatical recursion. (What are the humans and starlings doing, then?)

- But there was another limitation on the experimental data: only sentences up to a small fixed length (n=3) were presented. What would a statistical learner do in the context of such limited data, even if it weren't equiprobable?

Data restrictions

- I applied the inside-outside algorithm to the same grammar, removing all sentences from the stochastic training data of length greater than 6. This yielded the following grammar:

$$
\begin{aligned}
(0)\ \mathrm{S} &\rightarrow \mathrm{A\ S\ B} \\
(.471)\ \mathrm{S} &\rightarrow \mathrm{A\ B} \\
(.300)\ \mathrm{S} &\rightarrow \mathrm{A\ A\ B\ B} \\
(.229)\ \mathrm{S} &\rightarrow \mathrm{A\ A\ A\ B\ B\ B} \\
(0)\ \mathrm{S} &\rightarrow \mathrm{A\ A\ A\ A\ B\ B\ B\ B}
\end{aligned}
$$

- This result is optimal from the perspective of the statistical learner: it does not need to throw away probability mass on the possibility of generating strings longer than those actually witnessed in the training data. This tendency can be somewhat tempered by a Bayesian approach to grammar learning, but is difficult to overcome in the face of such limited data.

Inducing recursion

- Question: How can we overcome the desire of the learner not to throw away probability mass on unseen strings?
- Answer: inductive bias (universal grammar)
- Another Question: But what does this inductive bias look like?

The nature of inductive bias

- The Bayesian approach to learning:

Evaluate the probability of hypotheses given the data

$$p(H|D) \propto P(D|H)p(H)$$

posterior probability of the hypothesis

likelihood of the data

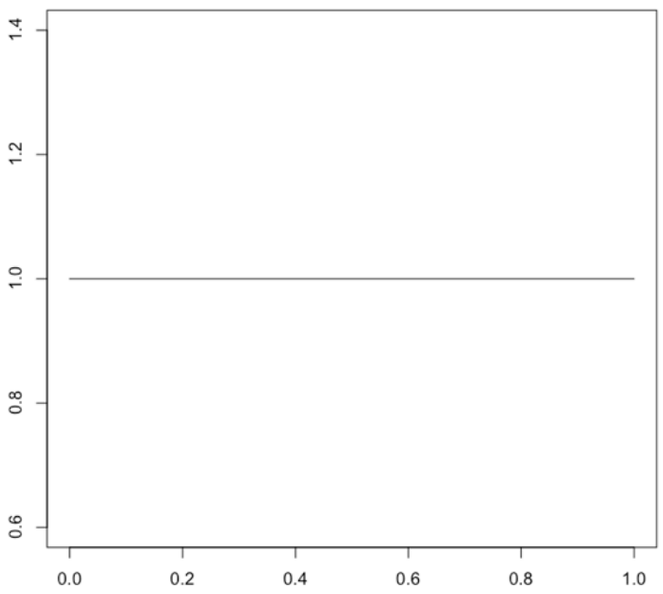
prior probability of the hypothesis

Bayesian inductive bias

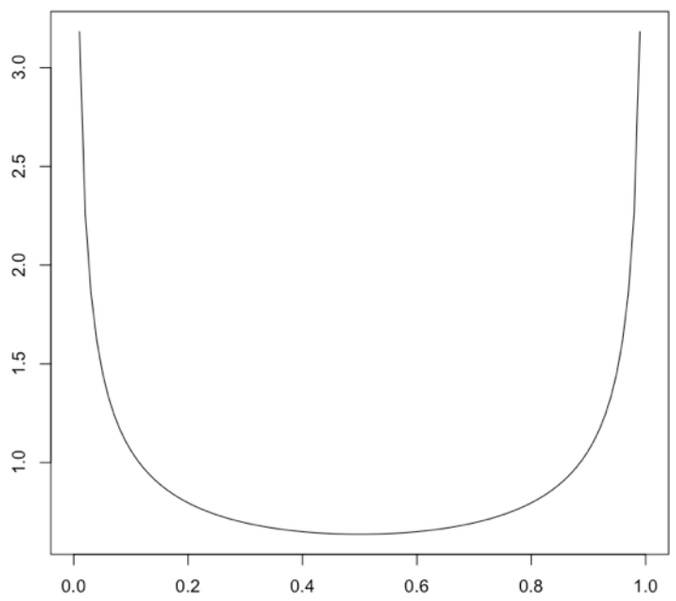
- What is the nature of the prior probability distribution? That is, how likely does a learner consider different grammatical hypotheses?

  - A proposal: prefer hypotheses with small or large probability values (i.e., penalize splitting rule probabilities among many rules)

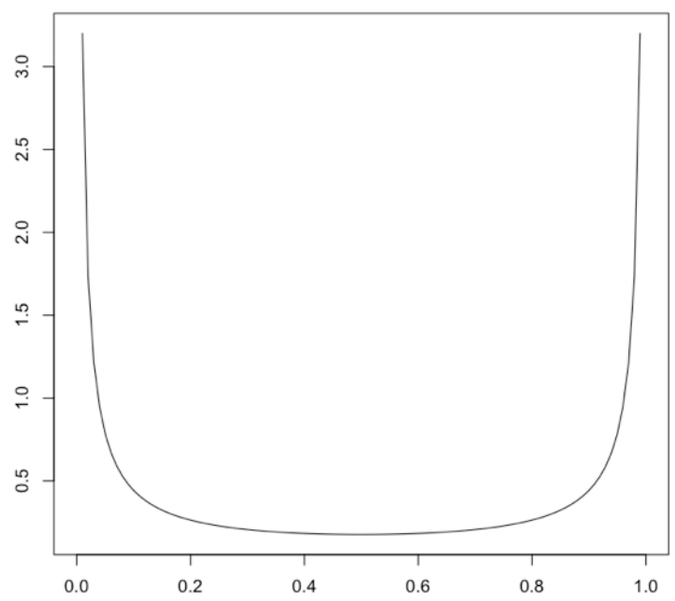
    Dirichlet prior: α parameter tells us how much we dislike spreading probability around.



α=1



α=.5



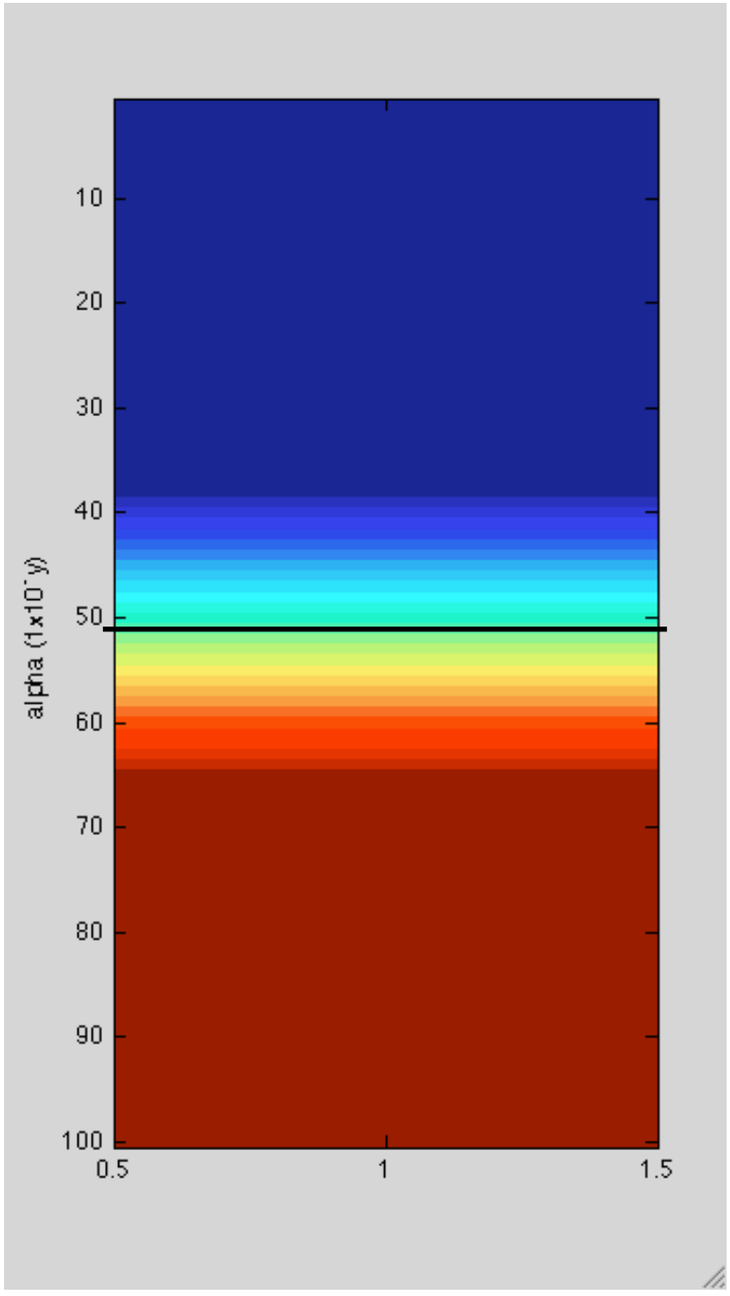
α=.1

Bayesian inductive bias

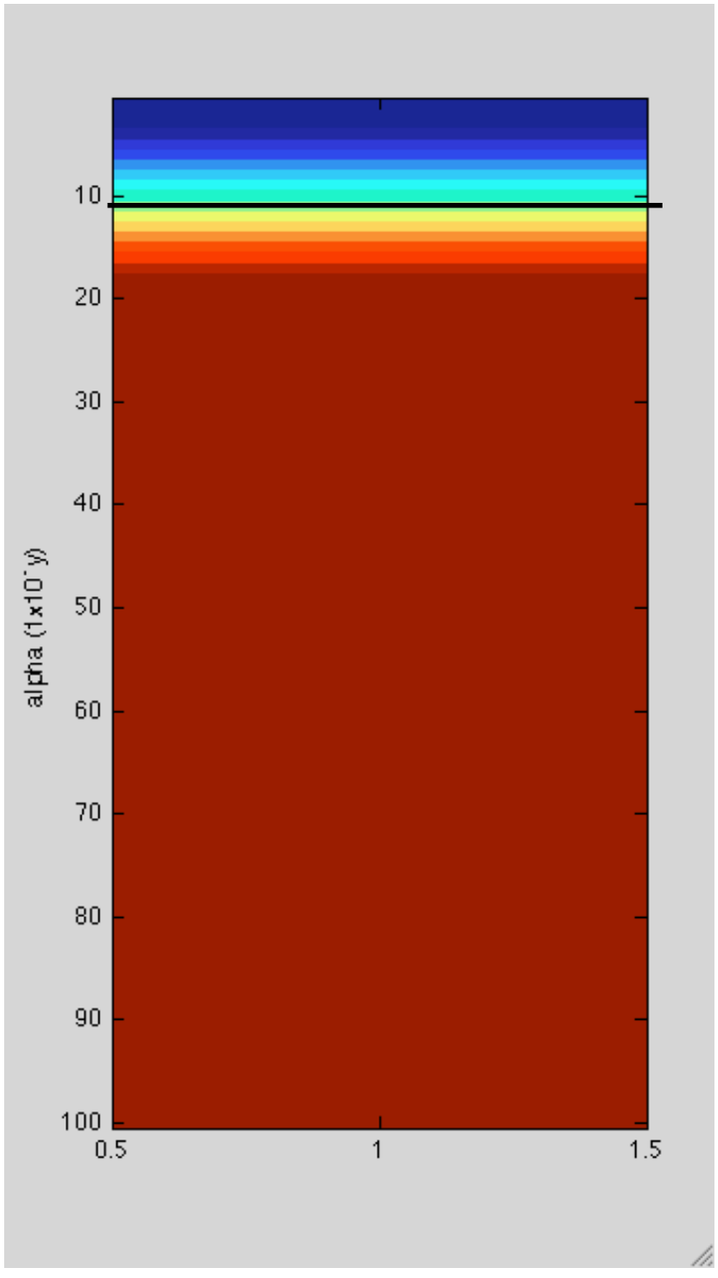
- An experiment:
  - Compare posterior probabilities of recursive and non-recursive hypothesis given
    - different strengths of prior (i.e., alpha values)
    - different data samples (i.e., stochastic vs. flat distributions)

Bayesian inductive bias

flat distribution



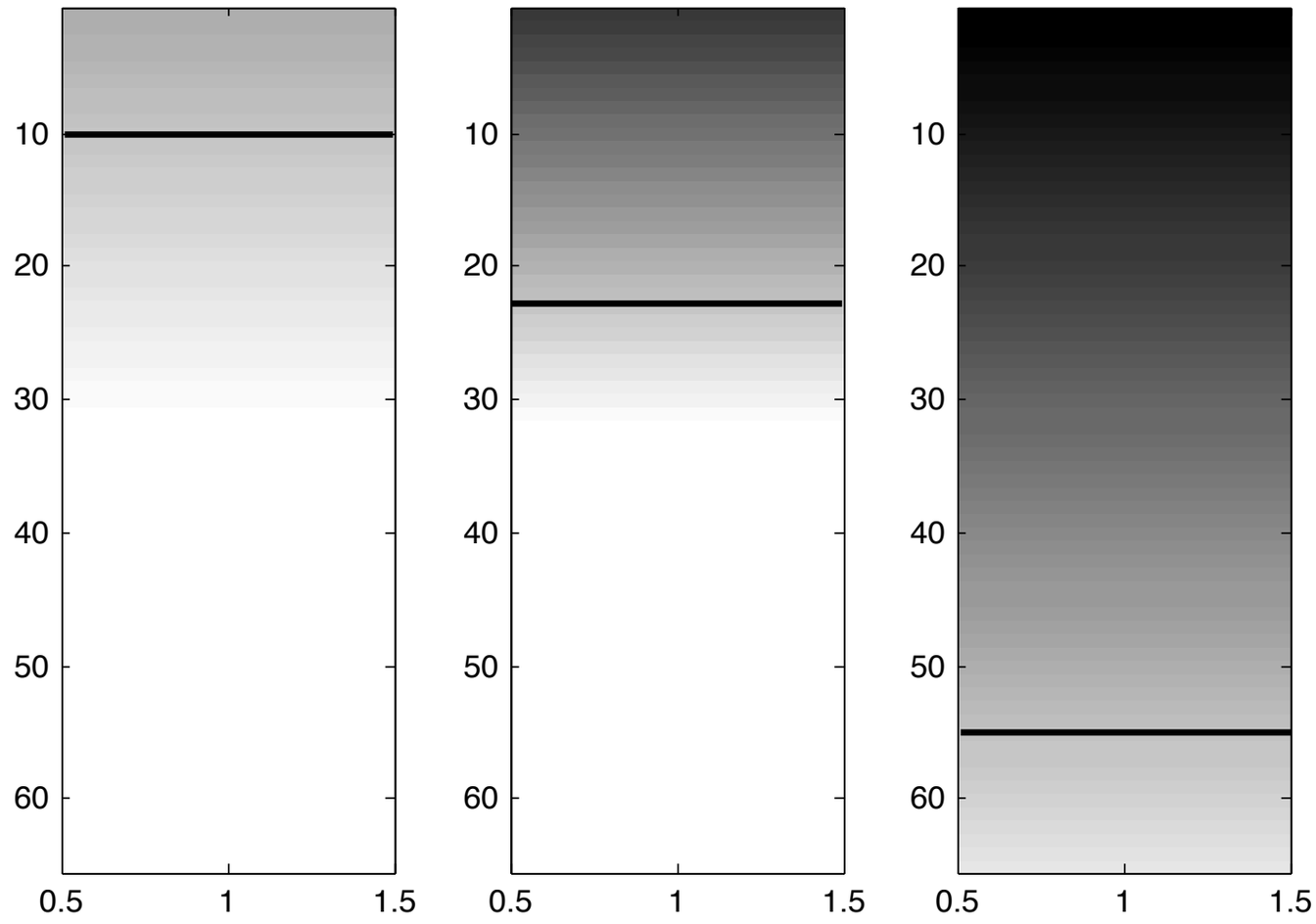
stochastic distribution



Color indicates ratio of posterior probability for recursive vs. non-recursive hypotheses (black line =1)

Bayesian inductive bias

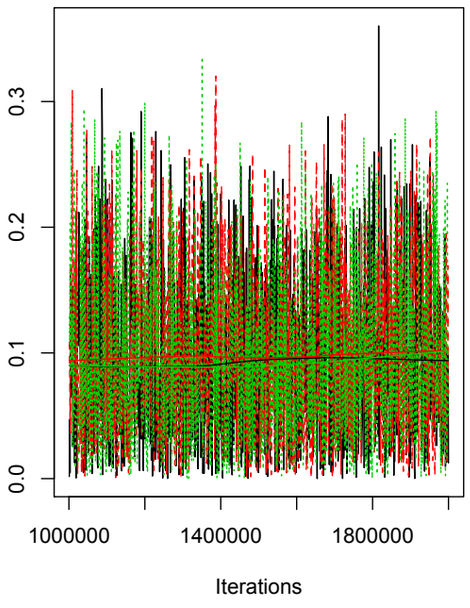
- Size of data sample also matters: as stochastic sample increases in size, it becomes *more difficult* to get the learner to conclude that there is a recursive grammar

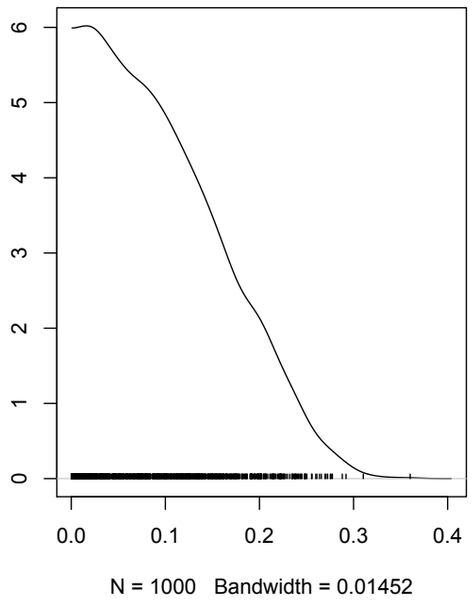


Bayesian Inductive Bias

- Thus far, we have only considered the relative goodness of two hypotheses: the completely recursive and the completely non-recursive grammars. Might an intermediate grammar have an even higher posterior probability?

- Use MCMC methods (Gibbs Sampling) to estimate posterior probability of recursion given stochastically distributed data (*k* examples of A^nB^n, with $n \in [1,5]$, $p_{rec}=.3$)





α=1
k=100

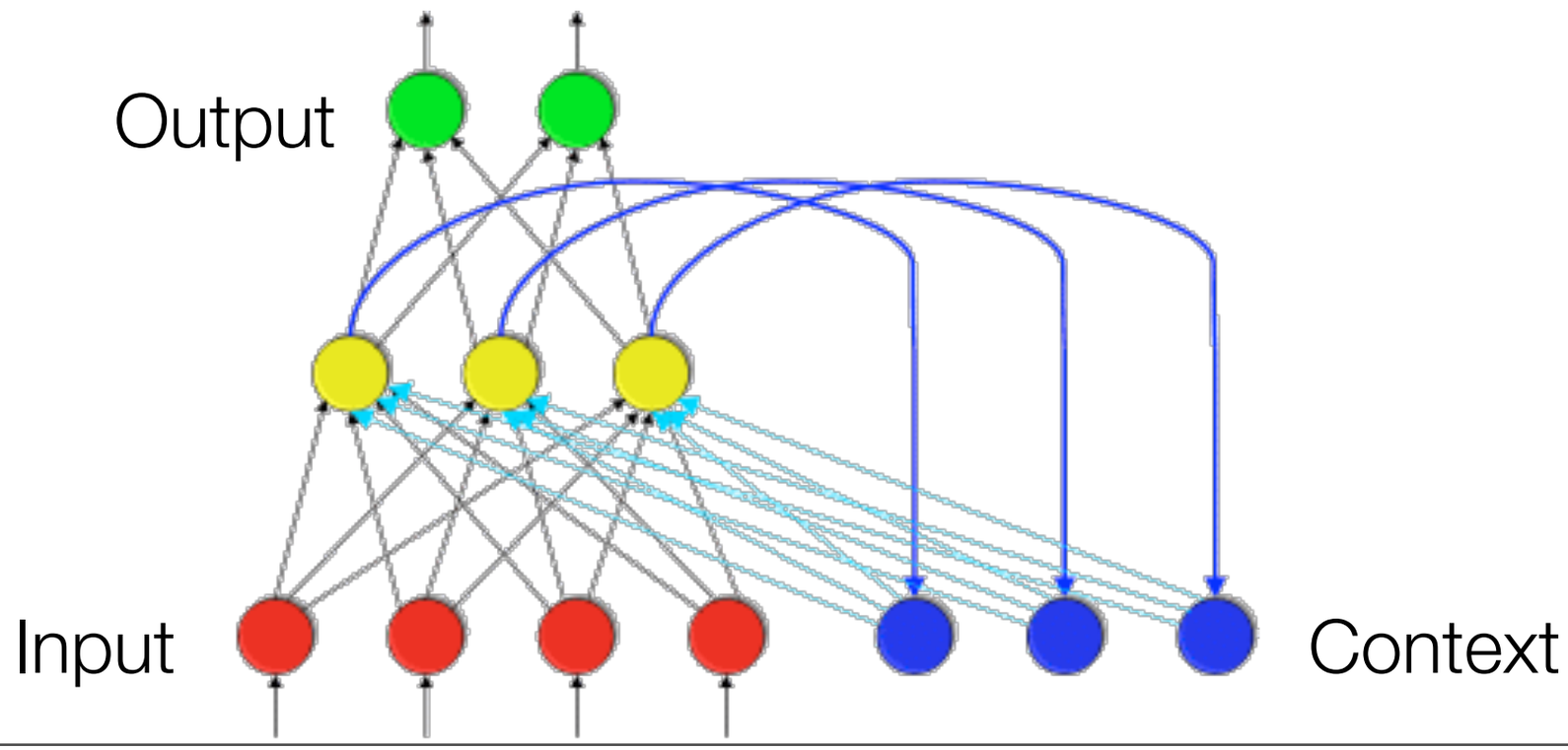
mean =.09

	α=1	α=.5	α=.1
k=100	0.09	0.11	0.25
k=1000	0.12	0.25	0.30

Another kind of inductive bias: connectionist networks

- Elman (1990,1991): represent sequences through temporal extent
  - Feed a sequence to a connectionist network one symbol at a time
  - Activation of hidden units is copied to **context units** at each time step
  - Context units provide input to hidden units at the next time, providing memory of past.
  - Identical inputs can be treated differently, depending on context.

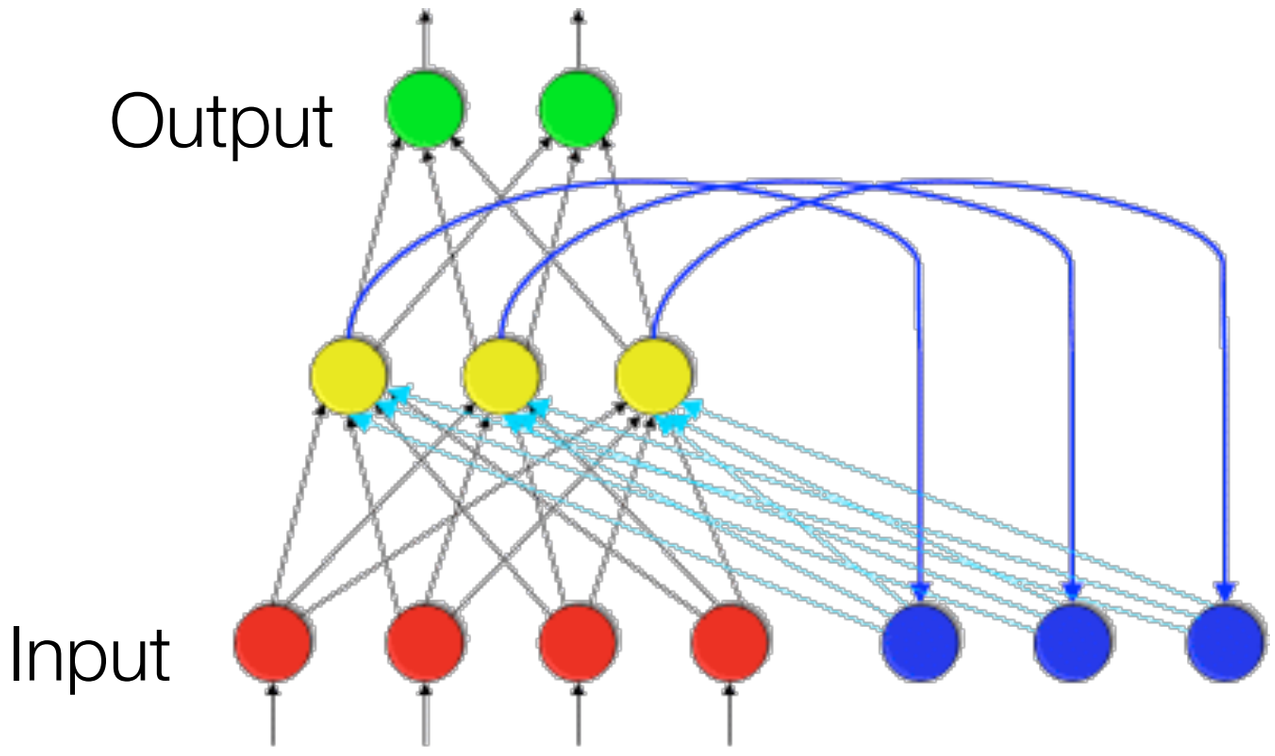
Simple Recurrent Network



Connectionist networks

- Train network to predict the next symbol in the sequence, adjusting the weights between the units when the output is inaccurate.

- Though it is in general impossible to accurately predict the next symbol, knowledge of the structure of a language can help to restrict the possible guesses.



Where's the bias?

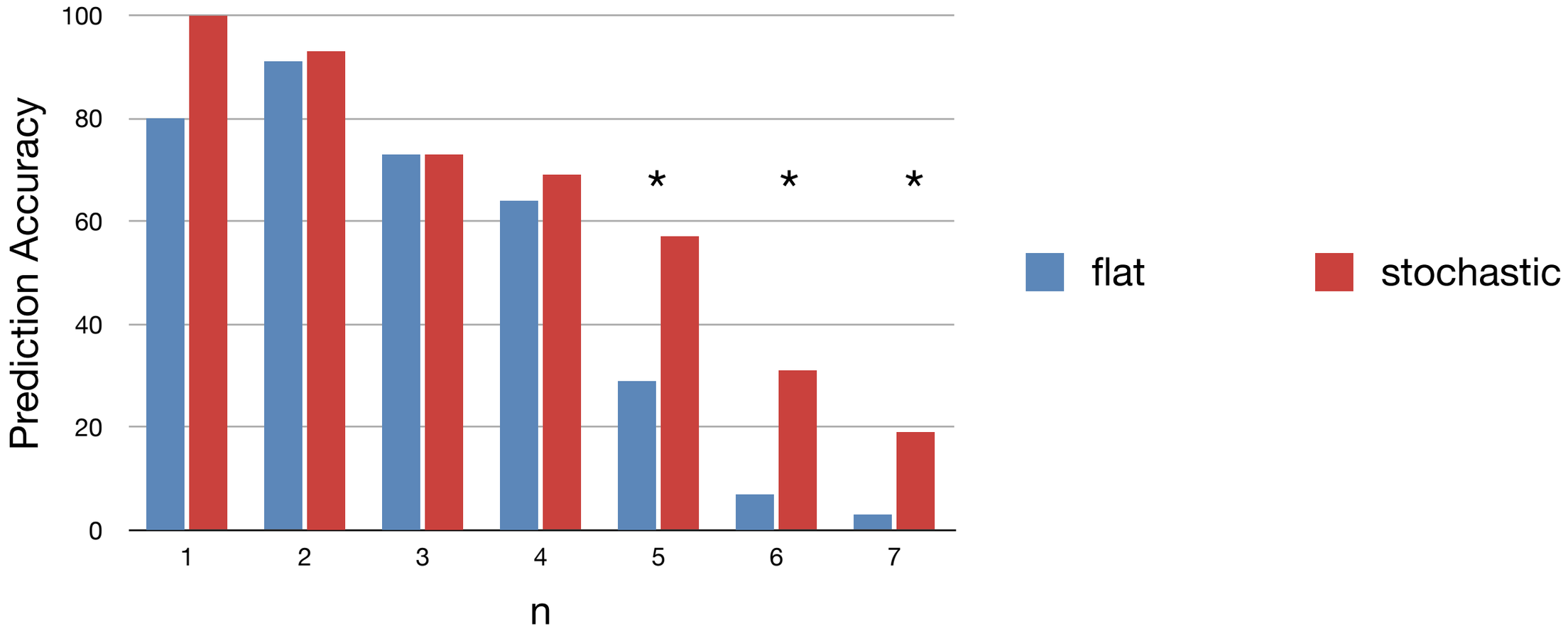
- Number of hidden units (compression of knowledge)

- Properties of learning algorithm

Another kind of inductive bias: Connectionist networks

- Trained networks to predict next symbol and end of sentence symbol for stochastically generated and flatly distributed training set for strings from $a^n b^n$ language up to length 6 (i.e., no longer than $a^3 b^3$).

- Assessed accuracy of prediction of end of string on novel test set for strings of different lengths.



Conclusions and Implications

- Statistical distributions *need not* be orthogonal to structural properties of language: grammatical structure will typically have a statistical signature.

Statistical Signatures of Grammatical Structure

- Guy (1991): Phonological variation and t/d-deletion

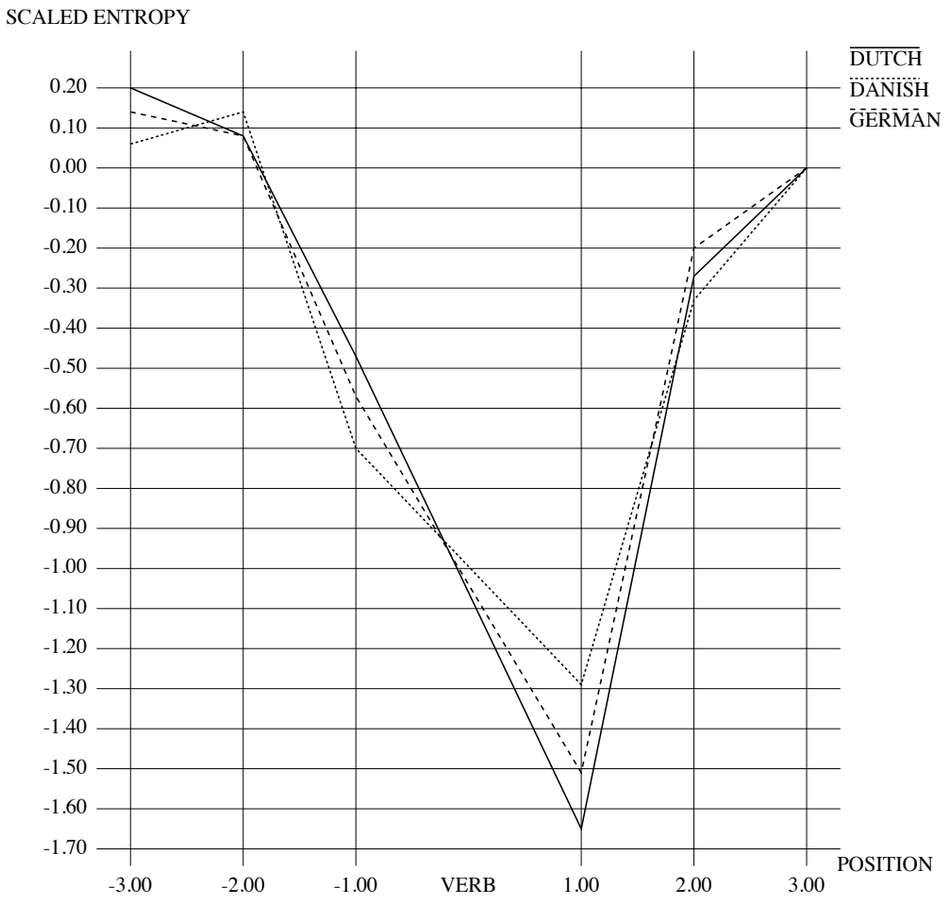
TABLE 1. *The exponential relationship:* -t,d *retention in two data sets*

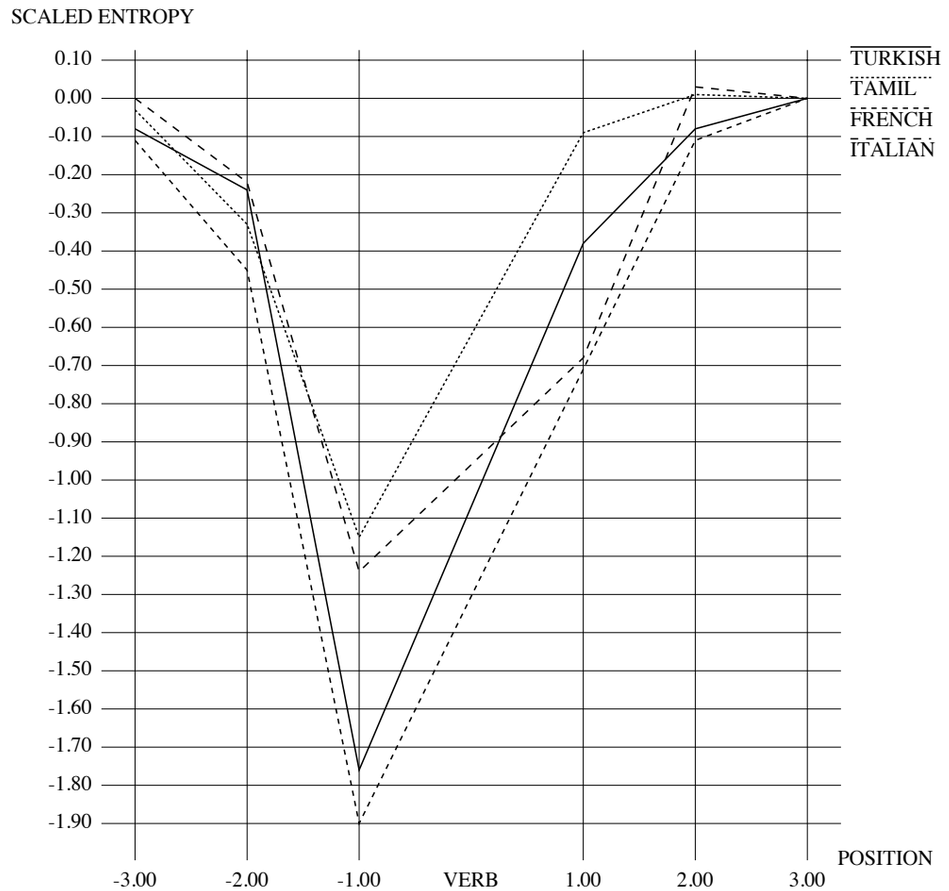
	N	% Ret.	Estimated p_r	Significance
Guy (1991) (7 speakers)				
Monomorphemic	658	61.9	.852	Best-fit $p_r = .85$
Semiweak past	56	66.1	.813	Chi-square = 1.28, $p = .55$
Regular past	181	84.0	.840	
Santa Ana (1991) (45 speakers)				
Monomorphemic	3724	42.1	.7494	Best-fit $p_r = .75$
Semiweak past	297	59.3	.7698	Chi-square = 1.17, $p = .57$
Regular past	836	74.3	.7428	

Statistical Signatures of Grammatical Structure

- Brill and Kapur (1993): Verb second and conditional entropy

$$-\sum_{w}\sum_{v} p_{+1}(w,v)\log p_{+1}(w|v)$$





Conclusions and Implications

- Statistical distributions *need not* be orthogonal to structural properties of language: grammatical structure will typically have a statistical signature.

- Such statistical signatures will also allow us to distinguish other types of grammatical hypotheses on the basis of the distributions they give rise, even below the level of context-free.

- A rational learner will attend to properties of statistical distributions while engaging in language acquisition.

- If we are to understand whether humans and non-humans possess the capacity to learn the patterns of natural and non-natural languages, we must conduct experiments where it is rational for them to draw such conclusions.