One model for the learning of language

Steve Piantadosi & Yuan Yang
UC Berkeley
Psychology & Neuroscience
Outline

- Overview of learnability & formal languages
- Learning model
  - Simple formal languages
  - Artificial language learning
  - Simplified English CFG
- Three related lines of ongoing work
  - Human experiments
  - Recursion in monkeys and human groups
  - Algorithm learning in indigenous Amazonians
Gold’s learnability result

- **Gold (1967)** showed that positive evidence is not enough for learners to necessarily identify their parent’s target grammar (see Johnson 2004)

```
L_x ⊂ L_y ⊂ L_1 ⊂ L_2 ⊂ L_3 ⊂ ... ⊂ L_∞
```

- **Gold’s theorem motivated a lot of theorizing about linguistic nativism**, averaging a citation every 4 days since 1967! (e.g. Wexler & Culicover’s *Formal Principles of Language Acquisition*)
6.3 The Logical Problem of Language Acquisition

What follows is a fairly technical proof of the idea that parts of our linguistic system are at least plausibly construed as an innate, in-built system. If you aren’t interested in this proof (and the problems with it), then you can reasonably skip ahead to section 6.4.

The argument in this section is that a productive system like the rules of Language probably could not be learned or acquired. Infinite systems are in principle, given certain assumptions, both unlearnable and unacquirable. Since we’ll show that syntax is an infinite system, we shouldn’t have been able to acquire it. So it follows that it is built in. The argument presented here is based on an unpublished paper by Alec Marantz, but is based on an argument dating back to at least Chomsky (1965).

First here’s a sketch of the proof, which takes the classical form of an argument by modus ponens:

Premise (i): Syntax is a productive, recursive and infinite system.  
Premise (ii): Rule-governed infinite systems are unacquirable.  
Conclusion: Therefore syntax is an unacquirable system. Since we have such a system, it follows that at least parts of syntax are innate.

The so-called Innateness Hypothesis, which claims that crucial components of our tacit linguistic knowledge are not learned through experience but are given by our biological/genetic specifications, is not really a hypothesis. Rather, it is an empirical conclusion mainly based on observations of child language acquisition, one of which is now known as the Argument from the Poverty of Stimulus (APS).
Child-directed speech supports hierarchical structure.

Child-directed speech would lead an ideal learner to choose a hierarchical grammar over alternatives.
Child-directed speech supports hierarchical structure

Log prior, likelihood, and posterior probabilities of each hand-designed grammar for each level of evidence. Because numbers are negative, smaller absolute values correspond to higher probability. If two grammars have log probabilities that differ by \( n \), their actual probabilities differ by \( e^n \); thus, the best hierarchical phrase-structure grammar CFG-L is \( e^{191} (\sim 10^{53}) \) times more probable than the best linear grammar REG-M. Bold values indicate the highest posterior score at each level.

| Corpus | Probability | FLAT | REG-N | REG-M | REG-B | 1-ST | CFG-S | CFG-L |
|--------|-------------|------|-------|-------|-------|------|-------|-------|
| Level 1 | Prior       | -99  | -148  | -124  | -117  | -94  | -155  | -192  |
|         | Likelihood  | -17  | -20   | -19   | -21   | -36  | -27   | -27   |
|         | Posterior   | -116 | -168  | -143  | -138  | -130 | -182  | -219  |
| Level 2 | Prior       | -630 | -456  | -442  | -411  | -201 | -357  | -440  |
|         | Likelihood  | -134 | -147  | -157  | -162  | -275 | -194  | -177  |
|         | Posterior   | -764 | -603  | -599  | -573  | -476 | -551  | -617  |
| Level 3 | Prior       | -1198| -663  | -614  | -529  | -211 | -454  | -593  |
|         | Likelihood  | -282 | -323  | -333  | -346  | -553 | -402  | -377  |
|         | Posterior   | -1480| -986  | -947  | -875  | -764 | -856  | -970  |
| Level 4 | Prior       | -5839| -1550 | -1134 | -850  | -234 | -652  | -1011 |
|         | Likelihood  | -1498| -1761 | -1918 | -2042 | -3104| -2078 | -1956 |
|         | Posterior   | -7337| -3111 | -3052 | -2892 | -3338| -2730 | -2967 |
| Level 5 | Prior       | -10,610| -1962 | -1321 | -956  | -244 | -732  | -1228 |
|         | Likelihood  | -2856| -3376 | -3584 | -3816 | -5790| -3917 | -3703 |
|         | Posterior   | -13,466| -5338 | -4905 | -4772 | -6034| -4649 | -4931 |
| Level 6 | Prior       | -67,612| -5231 | -2083 | -1390 | -257 | -827  | -1567 |
|         | Likelihood  | -18,118| -24,454| -25,696| -27,123| -40,108| -27,312| -26,111 |
|         | Posterior   | -85,730| -29,685| -27,779| -28,513| -40,365| -28,139| -27,678 |
More optimistic results about positive evidence

- Positive evidence can lead you to the correct answer out of all computations (Chater & Vitanyi, 2007).
'Ideal learning' of natural language: Positive results about learning from positive evidence

Nick Chatera,b, Paul Vitanyib

aDepartment of Psychology, University College, London WC1E 6BT, UK
bUniversity of Waterloo & Infrasense, Waterloo, The Netherlands

1 Introduction

Language acquisition involves the ability of the linguistic structure of an unknown language to be learned from noisy and partial input. One influential line of argument is that the "poverty of the stimulus" argument (CI) is insufficient to account for the ability of children to learn language. However, a more recent line of argument is that children are able to learn language from a limited amount of input. In this paper, we present evidence that children are able to learn language from a limited amount of input.

In Part I, four ostensibly different models are presented, in which the probability of a very long sequence of symbols is used to predict the model. In Part II, the models are extended to predict the model's output. Some strong heuristic arguments for the usefulness of these models are presented. In Part III, the Bayes formulation, in which a priori and posteriori probabilities are used to predict the model's output, is presented.

Universal Artificial Intelligence

Sequential Decisions Based on Algorithmic Probability

Marcus Hutter

R. J. Solomonoff

Rockford Research Institute, Inc., Cambridge, Massachusetts

Available online at www.sciencedirect.com

ScienceDirect

Journal of Mathematical Psychology 51 (2007) 156–198

INFORMATION AND CONTROL, vol. 7, 1–22 (1964)

Universal Artificial Intelligence

Sequential Decisions Based on Algorithmic Probability

Marcus Hutter

R. J. Solomonoff

Rockford Research Institute, Inc., Cambridge, Massachusetts

Available online at www.sciencedirect.com

ScienceDirect

Journal of Mathematical Psychology 51 (2007) 156–198

INFORMATION AND CONTROL, vol. 7, 1–22 (1964)
Outline

• Overview of learnability & formal languages

• Learning model
  • Simple formal languages
  • Artificial language learning
  • Simplified English CFG

• Three related lines of ongoing work
  • Human experiments
  • Recursion in monkeys and human groups
  • Algorithm learning in indigenous Amazonians
TWO MODELS FOR THE DESCRIPTION OF LANGUAGE

Noam Chomsky

Department of Modern Languages and Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts

Abstract

We investigate several conceptions of linguistic structure to determine whether or not they can provide simple and "revealing" grammars that generate all of the sentences of English and only these. We find that no finite-state Markov process that produces symbols with transition from state to state can serve as an English grammar. Furthermore, the particular subclass of such processes that produce n-order statistical approximations to English do not come closer, with increasing n, to matching the output of an English grammar. We formalize the notions of "phrase structure" and show that this gives us a method for describing language which is essentially more powerful, though still representable as a rather elementary type of finite-state process. Nevertheless, it is successful only when limited to a small subset of simple sentences. We study the formal properties of a set of grammatical transformations that carry sentences with phrase structure into new sentences with derived phrase structure, showing that the transformations are processes of the same elementary type as phrase structure grammars; that the grammar of English is materially simplified if phrase structure description is limited to a kernel of simple sentences from which all other sentences are constructed by repeated transformations; and that this view of linguistic structure gives a certain insight into the use and understanding of language.

1. Introduction

There are two central problems in the descriptive study of language. One primary concern of the linguist is to discover simple and "revealing" grammars for natural languages. At the same time, by studying the properties of such successful grammars and clarifying the basic concepts that underlie them, he hopes to arrive at a general theory of linguistic structure. We shall examine certain features of these related inquiries.

The grammar of a language can be viewed as a theory of the structure of this language. Any adequate theory is based on a certain finite set of observations and, by establishing general laws stated in terms of certain hypothetical constructs, it attempts to account for these observations, to show how they are interrelated, and to predict an indefinite number of new phenomena. A mathematical theory has the additional property that predictions follow rigorously from the body of theory. Similarly, a grammar is based on a finite number of observed sentences (the language's corpus) and it "projects" this set to an infinite set of grammatical sentences by establishing general "laws" (grammatical rules) framed in terms of such hypothetical constructs as the particular phrases, words, phrases, and so on, of the language under analysis. A properly formulated grammar should determine unambiguously the set of grammatical sentences.

General linguistic theory can be viewed as a metatheory which is concerned with the problem of how to choose such a grammar in the case of each particular language on the basis of a finite corpus of sentences. In particular, it will consider and attempt to explain the relation between the set of grammatical sentences and the set of observed sentences. In other words, linguistic theory attempts to explain the ability of a speaker to produce and understand new sentences, and to reject as ungrammatical other new sentences, on the basis of his limited linguistic experience.

Suppose that for any language there are certain clear cases of grammatical sentences and certain clear cases of ungrammatical sentences, e.g., (1) and (2), respectively, in English.

(1) I ate a sandwich
(2) John ate a

In this case, we can test the adequacy of a proposed linguistic theory by determining, for each language, whether or not the clear cases are handled properly by the grammar constructed in accordance with this theory. For example, if a large corpus of English does not happen to contain either (1) or (2), we ask whether the grammar that is determined for this corpus will project the corpus to include (1) and exclude (2). Even though such clear cases may provide only a weak test of adequacy for the grammar of a given language taken in isolation, they provide a very strong test for any general linguistic theory and for the set of grammars to which it leads, since we insist that in the case of each language the clear cases be handled properly in a fixed and predetermined manner. We can take certain steps towards the construction of an operational characterization of "grammatical sentences" that will provide us with the clear cases required to set the task of linguistics significantly.
Formal and natural languages

• Patterns in natural language correspond to different formal languages & require distinct computational resources (see Jäger & Rogers 2012)
  
  ● $a^n$ – {a, aa, aaa, aaaa, ...}
    The tall, angry, young, giraffe ...

  ● $(ab)^n$ – {ab, abab, ababab, ...}
    Bring two boats, three cups, six accordions, ...

  ● $a^n b^n$ – {ab, aabb, aaabbb, ...}
    If Ted cried then John was sad then John is empathetic.

  ● $ww$ – {abcabc, accbaaccba, bbbb, ...}

  ● $a^n b^m c^n d^m$ – {abcd, abbcdd, aabccd, ...}

• These examples face all the problems we started with – subset problem, infinite productivity, gold non-learnability, etc.
Working hypothesis

• Learning operates over a Turing-complete space

• **Learning is like programming** – learners combine existing operations in new ways to form generative models of data

• **More input data** drives revision, improvement of programs, justifying additional complexity.
Cartoon of learning setup

\[
F_0(x) := \text{pair}(a, \text{if}(\text{flip}(0.7), \varepsilon, F_0(\varepsilon)))
\]
\[
F_1(x) := \text{if}(\text{empty}(x), \varepsilon, \text{pair}(\text{first}(x), \text{pair}(F_1(\text{rest}(x)), b)))
\]
\[
F_2(x) := F_1(F_0(\varepsilon))
\]

F_0(x) := \text{pair}(a, \text{if}(\text{flip}(1/3), b, \text{pair}(F_0(\varepsilon), b)))

\text{aabb, ab, ab, aaabbb}
Outline

• Overview of learnability & formal languages

• Learning model
  • Simple formal languages
  • Artificial language learning
  • Simplified English CFG

• Three related lines of ongoing work
  • Human experiments
  • Recursion in monkeys and human groups
  • Algorithm learning in indigenous Amazonians
Learning generative programs can be fast and easy
The model builds distinct generative models

Learning curve

Learned program

Structure

def F():
    if flip():
        "a"
    else:
        return F() + "a"

def F():
    if flip():
        "ab"
    else:
        return "a" + F() + "b"
Learning generative programs can be fast and easy

(approximated on top strings)
Measure
- Precision
- Recall
- F
- Memorized (F)
(approximated on top strings)
tupirogolabubidakupadotigolabubidakutupiropadoti
tupirogolabubidakupadotigolabubidakutupiropadoti
tupirogolabubidakupadotigolabubidakutupiropadoti
tupirogolabubidakupadotigolabubidakutupiropadoti
tupirogolabubidakupadotigolabubidakutupiropadoti
tupirogolabubidakupadotigolabubidakutupiropadoti

Reber
Morgan & Newport

\[ S \rightarrow AP \ BP \ (CP) \]
\[ AP \rightarrow a \ (D) \]
\[ BP \rightarrow CP \ f \ | \ e \]
\[ CP \rightarrow c \ (g) \]

Morgan, Meier, & Newport

\[ S \rightarrow AP \ BP \ (CP) \]
\[ AP \rightarrow a \ a \ (d) \]
\[ BP \rightarrow a \ CP \ f \ | \ u \ e \]
\[ CP \rightarrow i \ c \ (g). \]
Morgan & Newport
\[ S \rightarrow AP \ BP \ (CP) \]
\[ AP \rightarrow a \ (D) \]
\[ BP \rightarrow CP \ f \ | \ e \]
\[ CP \rightarrow c \ (g) \]
Morgan, Meier, & Newport
\[ S \rightarrow AP \ BP \ (CP) \]
\[ AP \rightarrow a \ a \ (d) \]
\[ BP \rightarrow a \ CP \ f \ | \ u \ e \]
\[ CP \rightarrow i \ c \ (g) \]
Moving toward natural language

- Let’s give the learning model data from a LING-101 CFG, including a few kinds of structures – linear dependencies, tail recursion in AP, recursion in S, PP, etc.

\[
S \rightarrow NP \ VP \\
NP \rightarrow n \mid d \ n \mid d \ AP \ n \mid NP \ PP \\
AP \rightarrow a \mid a \ AP \\
VP \rightarrow v \mid v \ NP \mid v \ t \ S \mid VP \ PP \\
PP \rightarrow p \ NP
\]
Learning is *much* more powerful than POS has claimed

- **Hierarchical structure**  
  (Perfors, Tenenbaum & Regier 2011)

- **Language identification**  
  (Chater & Vitanyi 2007, Yang & Piantadosi 2022)

- **Phonology textbook problems**  
  (Ellis 2020)

- **Compositional semantics**  
  (Kwiatkowski et al. 2010)

- **Island constraints**  
  (Wilcox, Futrell, Levy 2021)

- **Linguistic features, structural generalizations**  
  (Warstadt et al. 2020, Warstadt & Bowman 2020)

- **Binding theory / c-command (programs on trees)**  
  (Gorensten & Piantadosi, in prep)
Modeling summary

- Idealized learners can construct computational devices to generate key structures in natural language.

- Children don’t need language-specific representations or biases to solve the learnability problem (though other evidence might make us think those are real)

- Complex, structured generalizations, infinite productivity, fast learning of latent generative processes – all are natural tendencies of learning systems that work over computations.
Outline

- Overview of learnability & formal languages
- Learning model
  - Simple formal languages
  - Artificial language learning
  - Simplified English CFG
- Three related lines of ongoing work
  - Human experiments
  - Recursion in monkeys and human groups
  - Algorithm learning in indigenous Amazonians
#1 – Formal language experiments

Ced Zhang
#2 – Bias for recursion

(Ferrigno, Cheyette, Piantadosi, & Cantlon 2020)
#2 – Bias for recursion
(Ferrigno, Cheyette, Piantadosi, & Cantlon 2020)
#2 – Bias for recursion
(Ferrigno, Cheyette, Piantadosi, & Cantlon 2020)

![Graph showing bias for recursion. The y-axis represents the proportion of trials ranging from 0 to 1.00, with error bars. The x-axis represents US adults. The graph indicates a significant difference (*). The response structure includes three types: Center-embedded, Crossed, and Tail-embedded.]
#2 – Bias for recursion
(Ferrigno, Cheyette, Piantadosi, & Cantlon 2020)
#2 – Bias for recursion
(Ferrigno, Cheyette, Piantadosi, & Cantlon 2020)
#2 – Bias for recursion
(Ferrigno, Cheyette, Piantadosi, & Cantlon 2020)
Working hypothesis

- Learning operates over a Turing-complete space
- Learning is like programming – learners combine existing operations in new ways to form generative models of data
- More input data drives revision, improvement of programs, justifying additional complexity.
#3 – Program induction in child learners

Ben Pitt

**Basic tasks**
- **Sort**
  - Response bins
  - Source bin (clear)
- **Double**
  - Response bins
  - Source bin
- **Hitch**
  - Response bins
  - Source bins

**Advanced tasks**
- **Random Sort**
  - Response bins
  - Source bin (opaque)
- **Color Switch**
  - Response bins
  - Source bin for demo
  - Source bin for test
- **Extended Sort**
  - Four response bins
  - Source bin (4 colors)
#3 – Program induction in child learners
Thank you

Thanks to my coauthors & contributors on this and related work: Yuan (Leon) Yang, Ced Zhang, Josh Rule, Steve Ferrigno, Sam Cheyette, Jessica Cantlon, Ben Pitt, Charlene Gallardo, Bin Li

Grant support:
NSF DRL 2000759
NSF DRL 1760874
NIH 1R01HD085996

Code is all available in our lab’s program induction library, Fleet: https://github.com/piantado/Fleet/