Harvard CS 121 and CSCI E-207

Lecture 9: Regular Languages Wrap-Up, Context-Free Grammars

Salil Vadhan

October 2, 2012

Reading: Sipser, §2.1 (except Chomsky Normal Form).

Algorithmic questions about regular languages

Given X = a regular expression, DFA, or NFA,
how could you tell if:

- $x \in L(X)$, where x is some string?
- $L(X) = \emptyset$?
- $x \in L(X)$ but $x \notin L(Y)$?
- $L(X) = L(Y)$, where Y is another RE/FA?
- $L(X)$ is infinite?
- There are infinitely many strings that belong to both $L(X)$ and $L(Y)$?

Generalizations of FA

Can add:

- probabilistic transitions (like Markov chains)
- outputs at each state
- rewards at each state
- infinite state spaces

Often referred to as “state machines”.

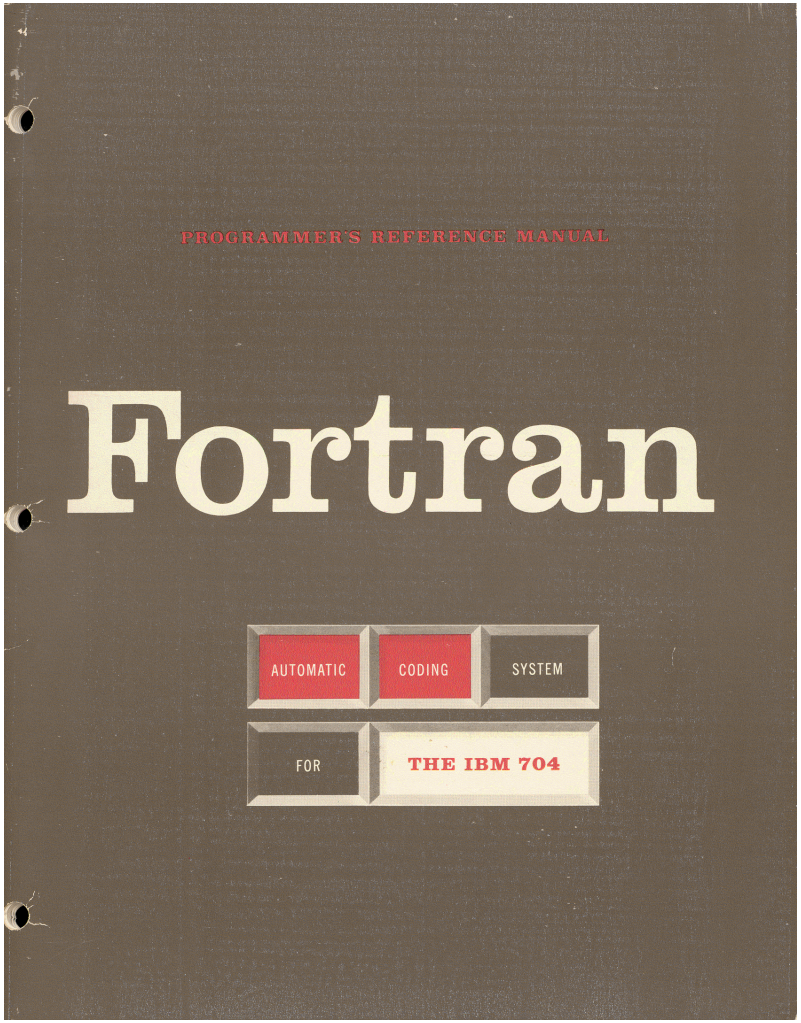
Applications of FA & generalizations

- pattern-matching algorithms (in software)
- control logic in CPUs and other hardware (input = instructions, outputs = control, one transition per clock cycle)
- low-level natural language processing (e.g. parts of speech, phonemes, morphemes but generally not syntax and grammar)
- components in distributed systems: I/O automata (inputs = from environment, outputs = to environment)
- decision making in AI, Econ, etc.: Markov Decision Processes (states = environment, inputs = actions of agent, outputs = observations of agent, rewards = payoff to agent).
- modelling in physical, biological, technological,... systems

FORTRAN

John Backus

The Fortran Automatic Coding System for the IBM 704 EDPM (October 15, 1956)



A Fortran Lexical Definition

Functions

As in the above example, a FORTRAN expression may include the name of a *function* (e.g. the sine function SINF), provided that the routine for evaluating the function is either built into FORTRAN or is accessible to it as a pre-written subroutine in 704 language on the master FORTRAN tape.

GENERAL FORM	EXAMPLES
The name of the function is 4 to 7 alphabetic or numeric characters (not special characters), of which the last must be F and the first must be alphabetic. Also, the first must be X if and only if the value of the function is to be fixed point. The name of the function is followed by parentheses enclosing the arguments (which may be expressions), separated by commas.	SINF(A+B) SOMEF(X,Y) SQRTF(SINF(A)) XTANF(3.*X)

A Fortran Syntactic Definition

Formal Rules for Forming Expressions. By repeated use of the following rules, all permissible expressions may be derived.

1. Any fixed point (floating point) constant, variable, or subscripted variable is an expression of the same mode. Thus 3 and I are fixed point expressions, and ALPHA and A(I,J,K) are floating point expressions.
2. If SOMEF is some function of n variables, and if E, F, ... , H are a set of n expressions of the correct modes for SOMEF, then SOMEF (E, F, ... , H) is an expression of the same mode as SOMEF.
3. If E is an expression, and if its first character is not + or −, then +E and −E are expressions of the same mode as E. Thus −A is an expression, but +−A is not.
4. If E is an expression, then (E) is an expression of the same mode as E. Thus (A), ((A)), (((A))), etc. are expressions.
5. If E and F are expressions of the same mode, and if the first character of F is not + or −, then

```
E + F
E - F
E * F
E / F
```

are expressions of the same mode. Thus A−+B and A/+B are not expressions. The characters +, −, *, and / denote addition, subtraction, multiplication, and division.

Peter Naur

Revised Report on the Algorithmic Language Algol 60 (1962)

1.1 Formalism for syntactic description.

The syntax will be described with the aid of metalinguistic formulae (1).

(1) Cf. J. W. Backus, The syntax and semantics of the proposed international algebraic language of the Zuerich ACM-GRAMM conference. ICIP Paris, June 1959.

Their interpretation is best explained by an example:

```
<ab> ::= ( | [ | <ab> ( | <ab> <d>
```

Sequences of characters enclosed in the bracket <> represent metalinguistic variables whose values are sequences of symbols. The marks ::= and | (the latter with the meaning of **or**) are metalinguistic connectives. Any mark in a formula, which is not a variable or a connective, denotes itself (or the class of marks which are similar to it). Juxta position of marks and/or variables in a formula signifies juxtaposition of the sequences denoted. Thus the formula above gives a recursive rule for the formation of values of the variable <ab>. It indicates that <ab> may have the value (or [or that given some legitimate value of <ab>, another may be formed by following it with the character (or by following it with some value of the variable <d>. If the values of <d> are the decimal digits, some values of <ab> are:

```
[(((1(37(
(12345(
(((
[86
```

Noam Chomsky

1956

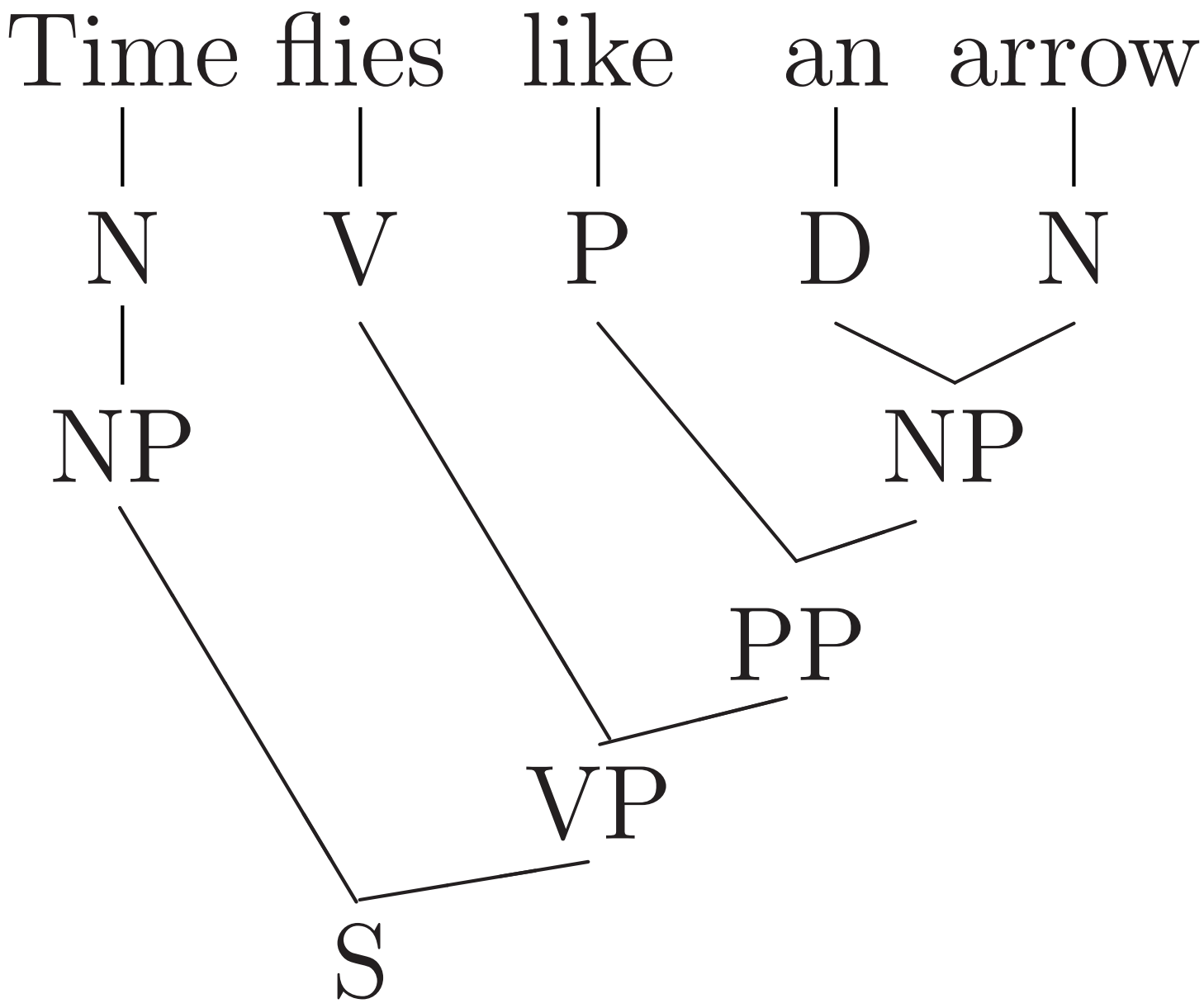
THREE MODELS FOR THE DESCRIPTION OF LANGUAGE*

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

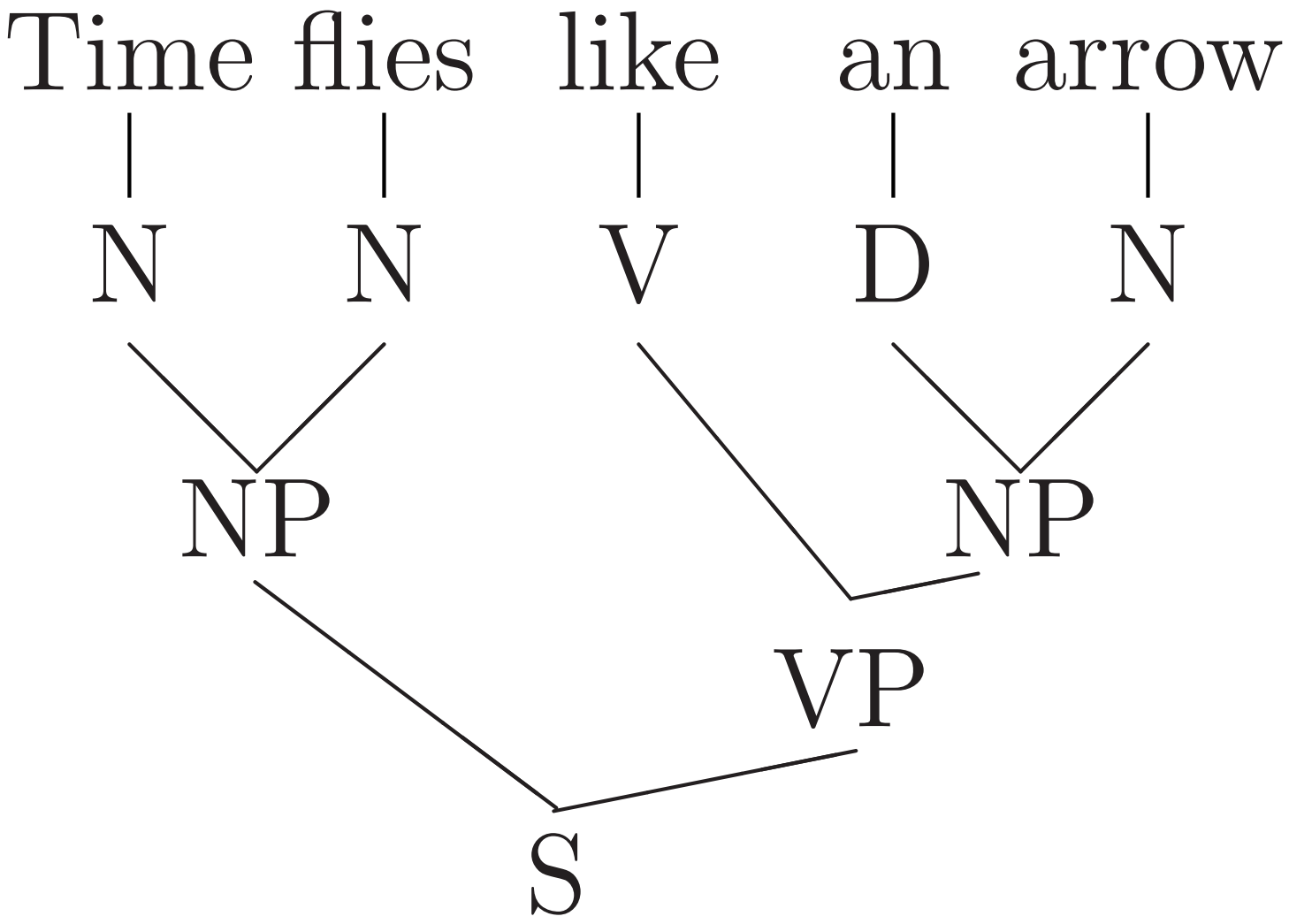
Abstract

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 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 linguist'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 phonemes, words, phrases, and so on, of the language under analysis. A properly formulated grammar should determine unambiguously the set of grammatical sentences.

Parse Trees



Parse Trees



Context-Free Grammars

- Originated as abstract model for:
  - Structure of natural languages (Chomsky)
  - Syntactic specification of programming languages (Backus-Naur Form)

- A context-free grammar is a set of generative rules for strings

  e.g.

$$G = \begin{array}{l} S \to aSb \\ S \to \varepsilon \end{array}$$

- A derivation looks like:

$$S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aabb$$

$$L(G) = \{\varepsilon, ab, aabb, \ldots\} = \{a^n b^n : n \geq 0\}$$

Equivalent Formalisms

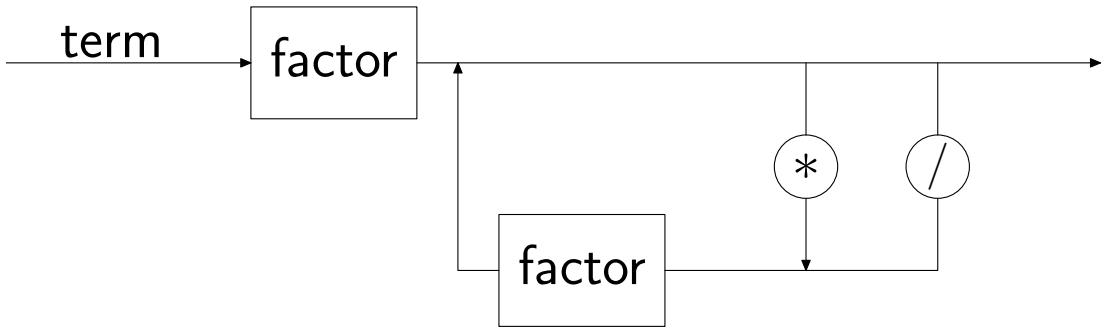
1. Backus-Naur Form (aka BNF, Backus Normal Form)

   due to John Backus and Peter Naur

   $$< \text{term} > ::= < \text{factor} > \mid < \text{factor} > * < \text{term} > \mid < \text{factor} > / < \text{term} >$$

   "$\mid$" means "or" in the metalanguage $=$ same left-hand side

2. "Railroad Diagrams"



Formal Definitions for CFGs

A CFG $G = (V, \Sigma, R, S)$

$V =$ Finite set of variables (or nonterminals)

$\Sigma =$ The alphabet, a finite set of terminals ($V \cap \Sigma = \emptyset$).

$R =$ A finite set of rules, each of the form $A \Rightarrow w$ for $A \in V$ and $w \in (V \cup \Sigma)^*$.

$S =$ The start variable, a member of V

e.g. $(\{S\}, \{a, b\}, \{S \to aSb, S \to \varepsilon\}, S)$

Derivations

For $\alpha, \beta \in (V \cup \Sigma)^*$ (strings of terminals and nonterminals),

$\alpha \Rightarrow_G \beta$ if $\alpha = uAv, \beta = uwv$ for some $u, v \in (V \cup \Sigma)^*$ and rule $A \rightarrow w$.

$\alpha \Rightarrow_G^* \beta$ ("α yields β") if there is a sequence $\alpha_0, \ldots, \alpha_k$ for $k \geq 0$ such that

$\alpha_0 = \alpha$, $\alpha_k = \beta$, and $\alpha_{i-1} \Rightarrow_G \alpha_i$ for each $i = 1, \ldots, k$.

$L(G) = \{w \in \Sigma^* : S \Rightarrow_G^* w\}$ (strings of terminals only!)

More examples of CFGs

- Arithmetic Expressions

G_1:

$$E \to x \mid y \mid E * E \mid E + E \mid (E)$$

G_2:

$$\begin{aligned} E &\to T \mid E + T \\ T &\to T * F \mid F \\ F &\to (E) \mid x \mid y \end{aligned}$$

Q: Which is “preferable”? Why?

More examples of CFGs, cont.

- $L = \{x \in \{(,)\}^* : \text{parentheses in } x \text{ are properly 'balanced'}\}$.

- $L = \{x \in \{a, b\}^* : x \text{ has the same \# of } a\text{'s and } b\text{'s}\}$.

Parse Trees

Derivations in a CFG can be represented by parse trees.

Examples:

Each parse tree corresponds to many derivations, but has unique leftmost derivation.

Parsing

Parsing: Given $x \in L(G)$, produce a parse tree for x. (Used to 'interpret' x. Compilers parse, rather than merely recognize, so they can assign semantics to expressions in the source language.)

Ambiguity: A grammar is ambiguous if some string has two parse trees.

Example: