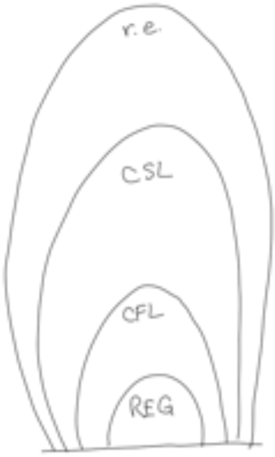
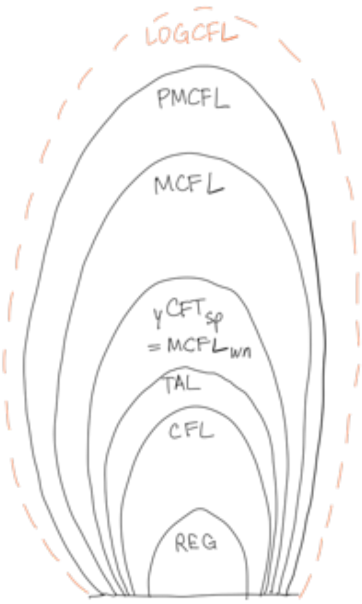
4/13 Mathematical Linguistics

Chomsky Hierarchy



Unrestricted Rewriting Systems $\alpha \to \beta$	Turing Machines
Context-Sensitive Grammars $\alpha \to \beta$ $\lvert\alpha\rvert \leq \lvert\beta\rvert$	Linear-Bounded Automata NSPACE(n)
Context-Free Grammars $A \to \beta$	Pushdown Automata
Right-Linear Grammars $A \to aB$ $A \to a$	Finite Automata
Grammars	Machines

More Important!!

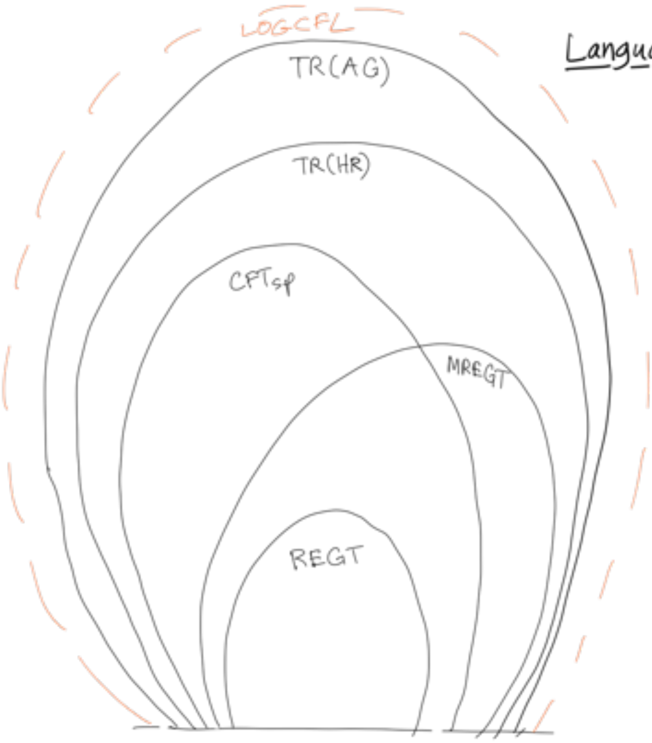


Parallel Multiple Context-Free Grammars

Multiple Context-Free Grammars

Simple Context-Free Tree Grammars
Well-Nested Multiple Context-Free Grammars

Tree-Adjoining Grammars



Languages of Trees

Attribute Grammars

Hyperedge Replacement Grammars

Simple Context-Free Tree Grammars

Multiple Regular Tree Grammars

Regular Tree Grammars

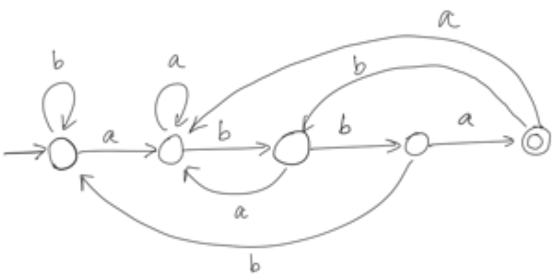
Why Formal Grammars?

- Mathematical theory of natural language
- Characterize "formal" properties of natural language
- Algorithmic models of
  - sentence processing
  - sentence production
  - grammar acquisition

But the fundamental reason for this inadequacy of traditional grammars is a more technical one. Although it was well understood that linguistic processes are in some sense "creative," the technical devices for expressing a system of recursive processes were simply not available until much more recently. In fact, a real understanding of how a language can (in Humboldt's words) "make infinite use of finite means" has developed only within the last thirty years, in the course of studies in the foundations of mathematics. Now that these insights are readily available it is possible to return to the problems that were raised, but not solved, in traditional linguistic theory, and to attempt an explicit formulation of the "creative" processes of language. There is, in short, no longer a technical barrier to the full-scale study of generative grammars.

Chomsky 1965, Aspects of the Theory of Syntax

Finite Automaton



$\{a,b\}^* abba$

DFA

$$\left(\underbrace{Q}_{\text{set of states}}, \Sigma, \delta, \underbrace{q_-}_{\text{initial state}}, \underbrace{F}_{\text{final states}}\right)$$

$$\delta : Q \times \Sigma \to Q$$

$$\hat{\delta} : Q \times \Sigma^* \to Q$$

$$\hat{\delta}(q, \varepsilon) = q$$

$$\hat{\delta}(q, wa) = \delta(\hat{\delta}(q, w), a)$$

$$\underbrace{\{w \in \Sigma^* \mid \hat{\delta}(q_-, w) \in F\}}_{\text{recognized language}}$$

Regular language

L is recognized by a DFA

$\iff$ $\simeq_L$ (Myhill-Nerode relation of L) has finite index

$\iff$ $\mathrm{Syn}(L)$ (syntactic monoid of L) is finite

$\iff$ L is recognized by a finite monoid

$$v \simeq_L w \underset{\text{def}}{\iff} \{u \in \Sigma^* \mid vu \in L\} = \{u \in \Sigma^* \mid wu \in L\}$$

syntactic congruence

$$v \equiv_L w \underset{\text{def}}{\iff} \{(x,y) \in \Sigma^* \times \Sigma^* \mid xvy \in L\} = \{(x,y) \in \Sigma^* \times \Sigma^* \mid xwy \in L\}$$

$$\mathrm{Syn}(L) = \Sigma^* / \equiv_L \quad \text{(quotient monoid)}$$

What are the equivalence classes of $\simeq_L$ and of $\equiv_L$?

- $L = \{a^n b^n \mid n \geq 0\}$ $\leftarrow$ not regular
- $L = \{a,b\}^* abba$

Context-Free Grammar

derivation tree

$S \to \varepsilon$

$S \to TS$

$T \to aSb$

$S \Rightarrow \underline{TS} \Rightarrow \underline{aSb}S \Rightarrow a\underline{TS}bS$

$\Rightarrow a\underline{aSb}SbS \Rightarrow aa\underline{\varepsilon}bSbS$

$\Rightarrow aa\varepsilon b\underline{\varepsilon}bS \Rightarrow aa\varepsilon b\varepsilon b\underline{TS}$

$\Rightarrow aa\varepsilon b\varepsilon b\underline{aSb}S \Rightarrow aa\varepsilon b\varepsilon ba\underline{\varepsilon}bS$

$\Rightarrow aa\varepsilon b\varepsilon ba\varepsilon b\underline{\varepsilon} = aabbab$

leftmost derivation

(())()

$aabbab \in L(G)$

Dyck language

productions

$$(N, \Sigma, P, S)$$

nonterminals

start symbol

A CFG G is self-embedding

$\iff$ for some nonterminal A and $x, y \in \Sigma^*$,

$$A \overset{*}{\underset{G}{\Rightarrow}} xAy \;\wedge$$

$$x \neq \varepsilon \;\wedge\; y \neq \varepsilon$$

Theorem (Chomsky) Let G be a CFG.

G is not self-embedding $\Rightarrow$ $L(G)$ is regular

Pumping Lemma

L is a CFL $\Rightarrow$

$$\exists p \in \mathbb{N}\ \forall z \in L$$
$$\Big(|z| \geq p \Rightarrow \exists uvwxy\ (z = uvwxy \;\wedge$$
$$|vx| \geq 1 \;\wedge$$
$$|vwx| \leq p \;\wedge$$
$$\forall n \in \mathbb{N}\ (uv^n w x^n y \in L)) \Big)$$

Show that the following languages are not context-free:

- $\{xx \mid x \in \{a,b\}^*\}$
- $\{a^n b^n c^n \mid n \geq 0\}$
- $\{a^m b^n a^m b^n \mid m, n \geq 0\}$

Swiss German

De Jan saït, dass mer (d'chind)^k (em Hans)^l es huus $\varphi_{k,l,m,n}$

said that we the children the house

haend wele laa^m hälfe^n aastriiche

have wanted let help paint

is grammatical if and only if $k = m \wedge l = n$

... that we have wanted to (let the children)^k (help Hans)^l paint the house

$$h\left(\text{Swiss German} \cap \{\varphi_{k,l,m,n} \mid k,l,m,n \geq 0\}\right) = \{a^m b^n a^m b^n \mid m,n \geq 0\}$$

h: d'chind $\mapsto a$ all other words $\mapsto \varepsilon$

Hans $\mapsto b$

laa $\mapsto a$

hälfe $\mapsto b$