Beyond CFGs

Computational Learning of Syntax

Alexander Clark

Department of Philosophy
King's College, London

LSA Summer Institute 2015, Chicago

Topics for today

- Some computational experiments on CFGs.
- Inadequacies of CFGs
- MCFGs as a representative MCS formalism
- Generalizing the notion of context
- Learning results
- Copying
- Learning copying

Overview

Maintain a set of contexts F and a set of strings K: construct a grammar $\mathcal{G}(K, L, F)$:

- Define nonterminals using small sets of contexts.
- Use strings to eliminate incorrect rules.

Overview

Maintain a set of contexts F and a set of strings K: construct a grammar $\mathcal{G}(K, L, F)$:

- Define nonterminals using small sets of contexts.
- Use strings to eliminate incorrect rules.

Monotonicity wrt contexts

If $F \subseteq G$ are two sets of contexts then
$\mathcal{L}(\mathcal{G}(K, L, F)) \subseteq \mathcal{L}(\mathcal{G}(K, L, G))$

Anti-Monotonicity with strings

If $J \subseteq K$ then $\mathcal{L}(\mathcal{G}(J, L, F)) \supseteq \mathcal{L}(\mathcal{G}(K, L, F))$

Primal

Representation

Each nonterminal in the grammar is represented by a small ($\leq k$) set of strings:

- Small set of strings W
- defines a nonterminal $[\![W]\!]$
- which we want to generate all the strings in $W^{\triangleright\triangleleft}$

Primal

Representation

Each nonterminal in the grammar is represented by a small ($\leq k$) set of strings:

- Small set of strings W
- defines a nonterminal $[\![W]\!]$
- which we want to generate all the strings in $W^{\triangleright\triangleleft}$

The larger the set W, the larger the set of strings $W^{\triangleright\triangleleft}$.

Use contexts to eliminate incorrect rules.

Primal algorithms

$$[\![A]\!] \to [\![B]\!][\![C]\!]$$

$$A^{\triangleright\triangleleft} \supseteq B^{\triangleright\triangleleft} C^{\triangleright\triangleleft}$$

Primal algorithms

$$[\![A]\!] \to [\![B]\!][\![C]\!]$$

$$A^{\triangleright\triangleleft} \supseteq B^{\triangleright\triangleleft} C^{\triangleright\triangleleft}$$

Incorrect if there is some context $l \Box r \in A^{\triangleright}$ and strings $u \in B, v \in C$ such that

- $luvr$ is not in L,

Theorem

This algorithm learns all context-free languages that have the FCP:

- Efficiently
- Correctly
- Weakly
- Using MQs.

Theorem

This algorithm learns all context-free languages that have the FCP:

- Efficiently
- Correctly
- Weakly
- Using MQs.

Discussion

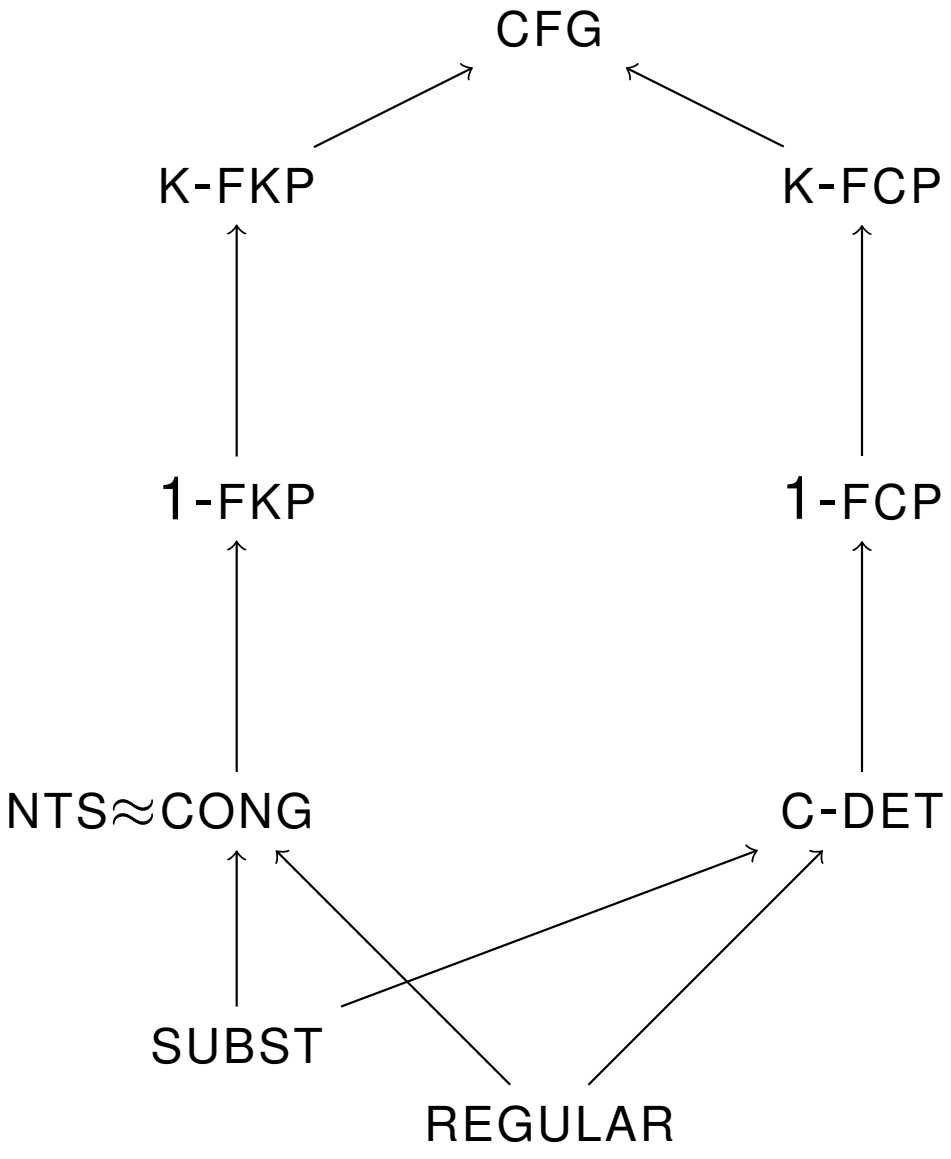
Is this really how children learn language?

Learnable languages

Weak learning, with MQs

- All finite languages
- All regular languages
- Not quite all context-free languages

Diagram



Unlearnable languages

Example

$L = \{a^n b^m \mid n \neq m\}$

$\{a, b, aab, abbb, aaaaabbb, \ldots\}$

Unlearnable languages

Example

$L = \{a^n b^m \mid n \neq m\}$

$\{a, b, aab, abbb, aaaaabbb, \ldots\}$

Why?

To represent this we need to use concepts that cannot be referred to by a finite number of contexts, or a finite number of strings.

Do these occur in natural language?

Inequality languages

$\{a^n b^m \mid n < m\}$

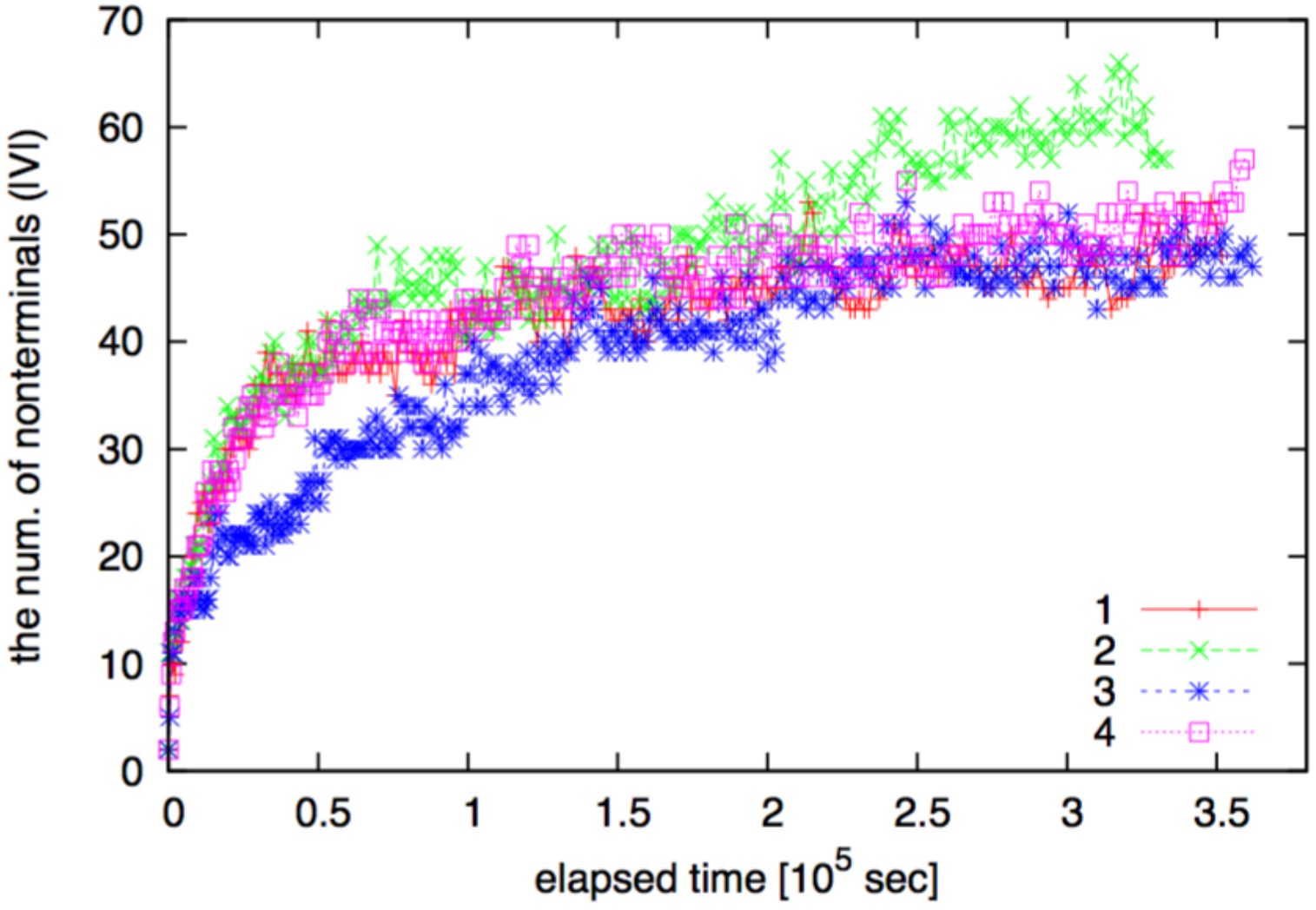
- (either) syntax or semantics
- (both) phonology and logic
- either either syntax or semantics or both phonology and logic
- either syntax or semantics or both phonology and logic
- syntax or semantics or both phonology and logic
- *either either syntax or semantics

Number of “either”s must be at most the number of “or”s.

Shibata 2014

- Non-parametric Bayesian model
- Finite context property.
- Tested on the Brown Corpus

Shibata, 2014



Shibata, 2014

Table 1: Comparing proposed methods with a baseline algorithm (modified Kneser-Ney). "Score" represents the average of log P(sentence) in the test data. The average of 4 trials is taken in the row of "$(1,0)$-context". Type-0 and type-2 nonterminals represent those which have a terminal and two nonterminals in the left-hand side of rules, respectively.

Method name	Score	The num. of type-0 nonterminals	The num. of type-2 nonterminals
$(0,0)$-context(blocked sampler)	27.043	60	37
$(0,0)$-context(proposed sampler)	25.775	122	94
$(1,0)$-context(proposed sampler)	**25.596**	46.0	6.75
modified Kneser-Ney(unigram)	39.407	-	-
modified Kneser-Ney(bigram)	27.067	-	-
modified Kneser-Ney(trigram)	25.802	-	-
modified Kneser-Ney(4-gram)	**25.675**	-	-
modified Kneser-Ney(5-gram)	25.823	-	-
modified Kneser-Ney(6-gram)	25.902	-	-

Scicluna, 2014

UC1: $a^n b^n$ **UC2**: $a^n b^n c^m d^m$ **UC3**: $a^n b^m\ n \geq m$ **UC4**: $a^p b^q$, $p \neq q$ **UC5**: Palindromes over alphabet $\{a, b\}$ with a central marker **UC6**:Palindromes over alphabet $\{a, b\}$ without a central marker **UC7**: Lukasiewicz language ($S \rightarrow aSS|b$)

and 4 described by ambiguous grammars:

AC1: $|w|_a = |w|_b$ **AC2**: $2|w|_a = |w|_b$ **AC3**: Dyck language **AC4**: Regular expressions.

The 9 artificial natural language grammars are:

NL1: Grammar 'a', Table 2 in (Langley and Stromsten, 2000) **NL2**: Grammar 'b', Table 2 in (Langley and Stromsten, 2000) **NL3**: Lexical categories and constituency, pg 96 in (Stolcke, 1994) **NL4**: Recursive embedding of constituents, pg 97 in (Stolcke, 1994) **NL5**: Agreement, pg 98 in (Stolcke, 1994) **NL6**: Singular/plural NPs and number agreement, pg 99 in (Stolcke, 1994) **NL7**: Experiment 3.1 grammar in (Adriaans et al., 2000) **NL8**: Grammar in Table 10 (Adriaans et al.,

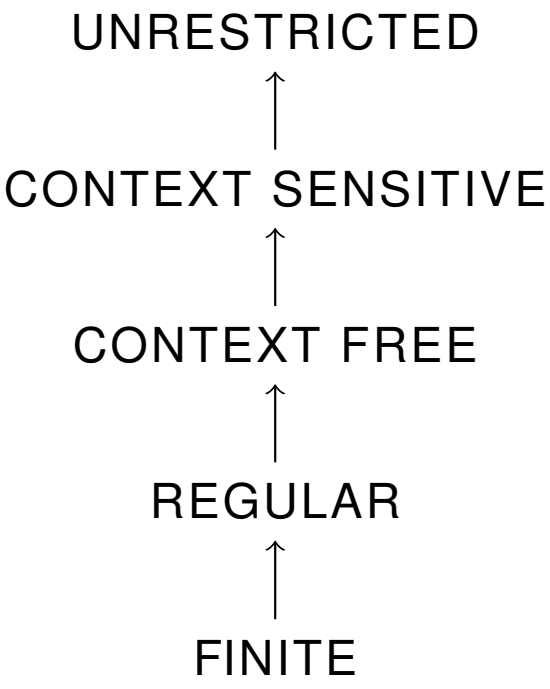
Scicluna, 2014

		Relative Entropy		UF_1	
Ex.	$\|S\|$	COMINO	ADIOS	COMINO	ABL
UC1	10	**0.029**	1.876	**100**	**100**
UC2	50	**0.0**	1.799	**100**	**100**
UC5	10	**0.111**	7.706	**100**	**100**
UC7	10	**0.014**	1.257	**100**	27.86
AC1	50	**0.014**	4.526	**52.36**	35.51
AC2	50	**0.098**	6.139	**46.95**	14.25
AC3	50	**0.057**	1.934	**99.74**	47.48
AC4	100	**0.124**	1.727	**83.63**	14.58
NL7	100	**0.0**	0.124	**100**	**100**
NL1	100	**0.202**	1.646	**24.08**	**24.38**
NL2	200	**0.333**	0.963	**45.90**	**45.80**
NL3	100	**0.227**	1.491	36.34	**75.95**
NL5	100	**0.111**	1.692	**88.15**	79.16
NL6	400	**0.227**	**0.138**	36.28	**100**
UC3	100	**0.411**	0.864	61.13	**100**
UC4	100	**0.872**	2.480	42.84	**100**
UC6	100	1.449	**1.0**	**20.14**	8.36
NL4	500	**1.886**	2.918	**65.88**	52.87
NL8	1000	**1.496**	**1.531**	**57.77**	50.04
NL9	800	1.701	**1.227**	12.49	**28.53**

Table 2: Relative Entropy and UF_1 results of our system COMINO vs ADIOS and ABL respectively. Best results are highlighted, close results (i.e. with a difference of at most 0.1 for relative

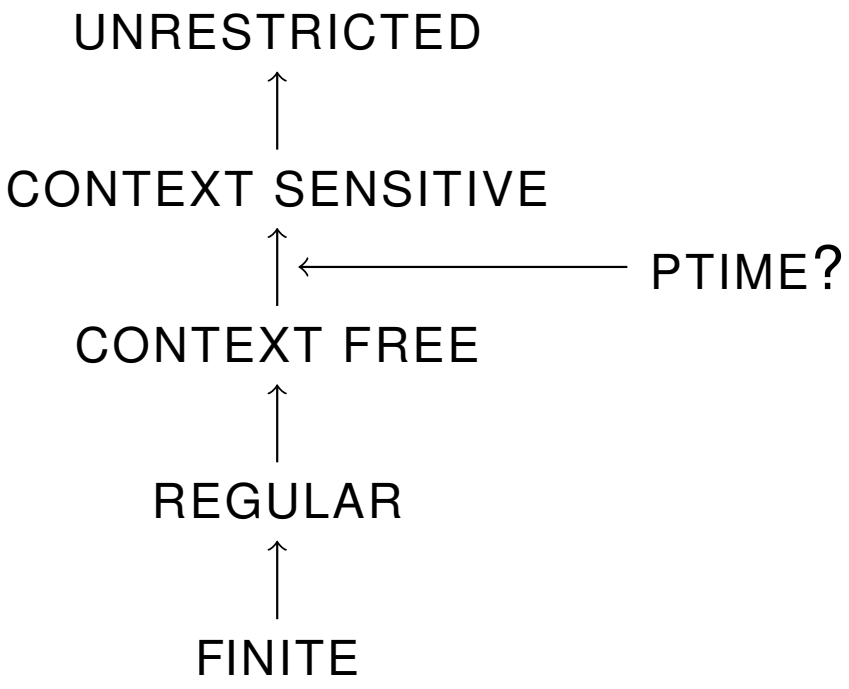
Chomsky hierarchy

Top down



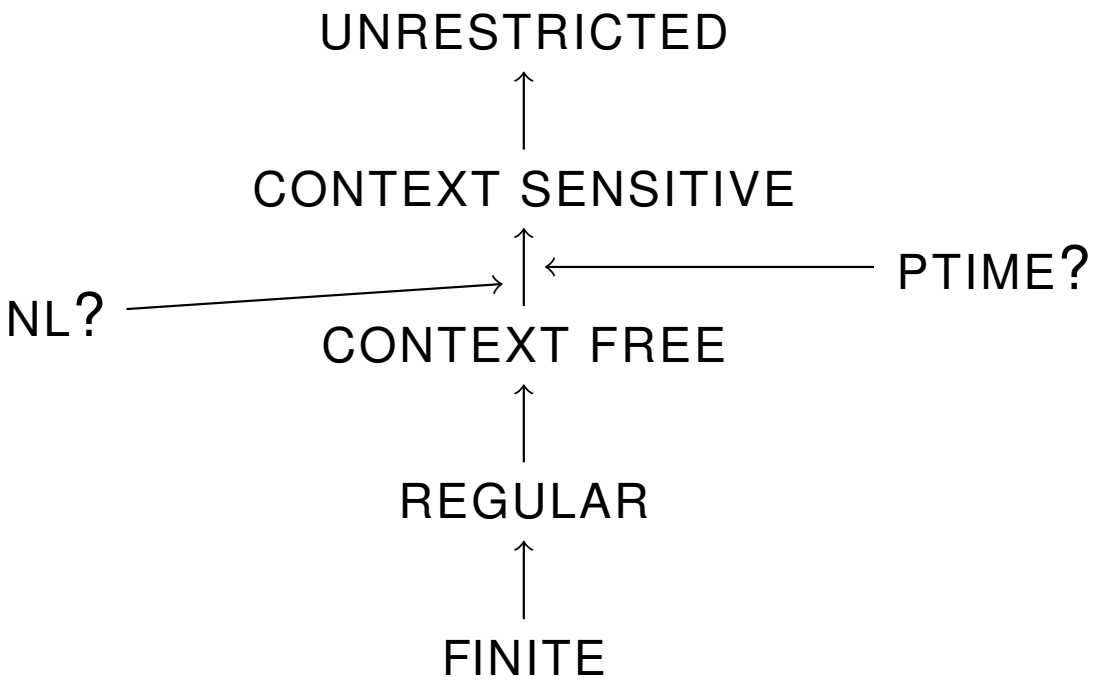
Chomsky hierarchy

Top down

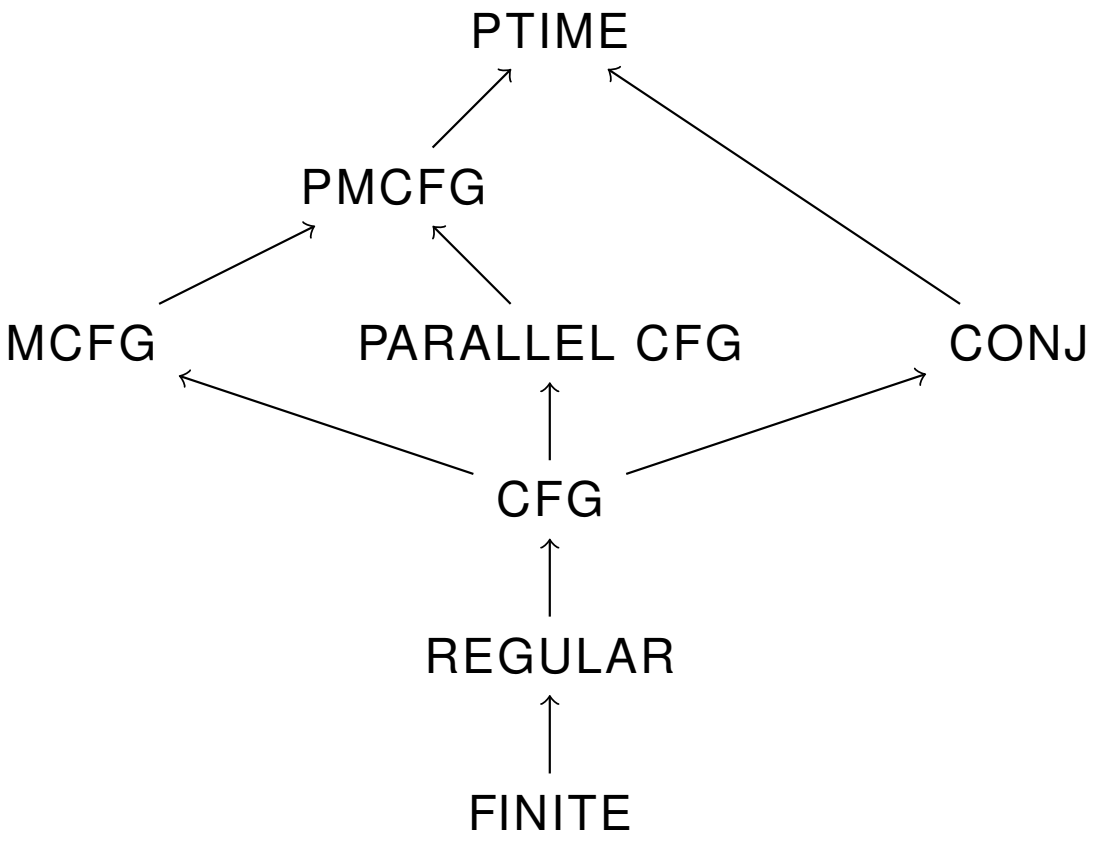


Chomsky hierarchy

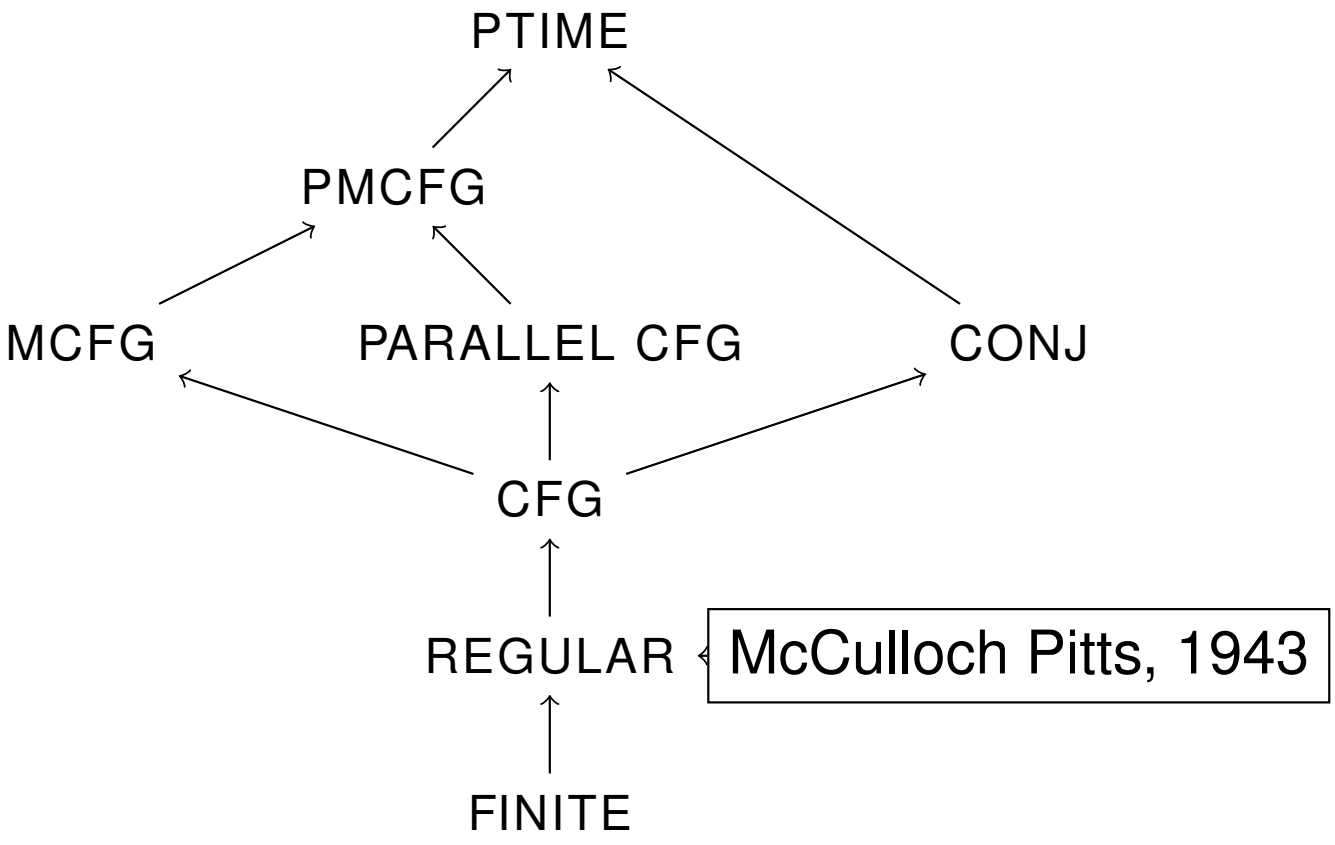
Top down



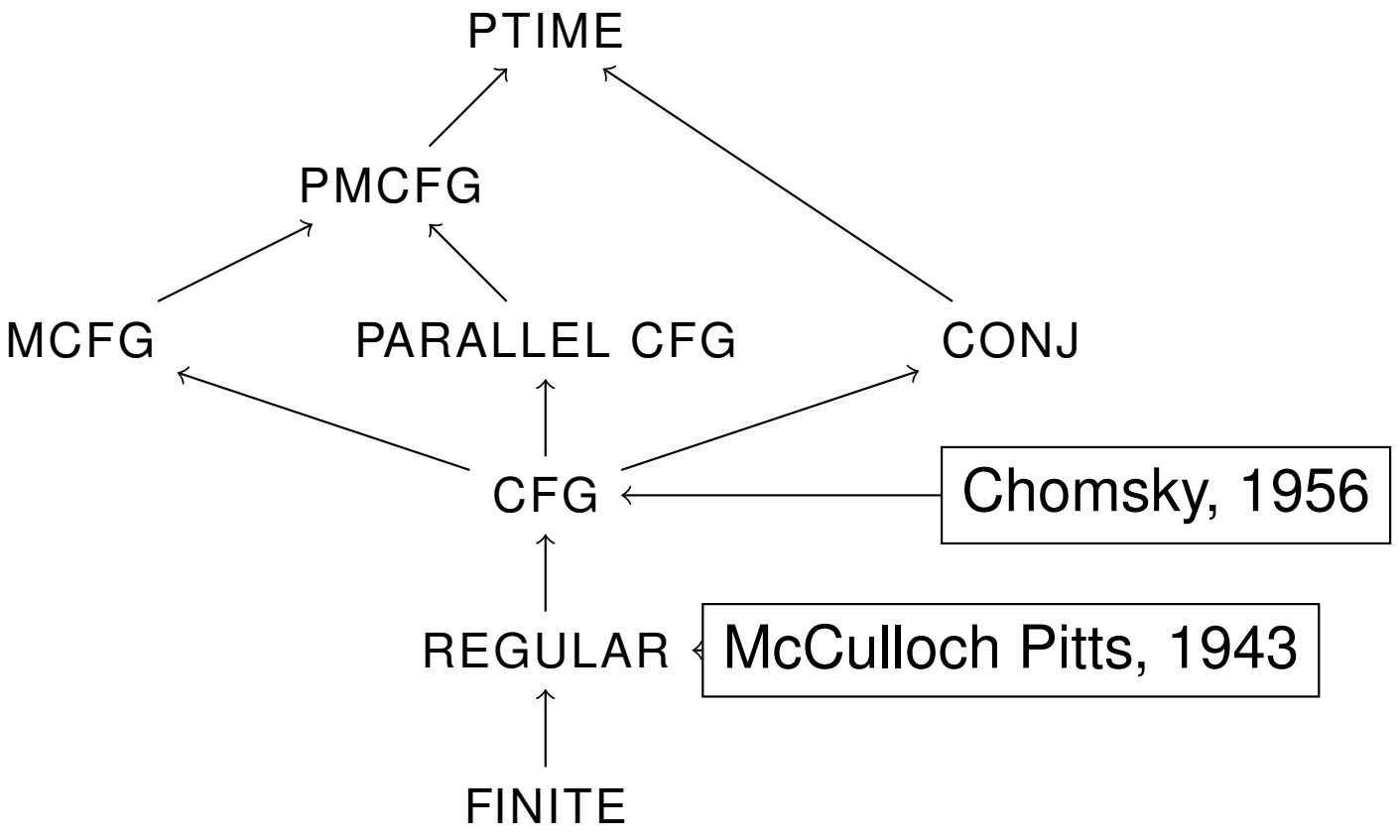
Grammar formalisms



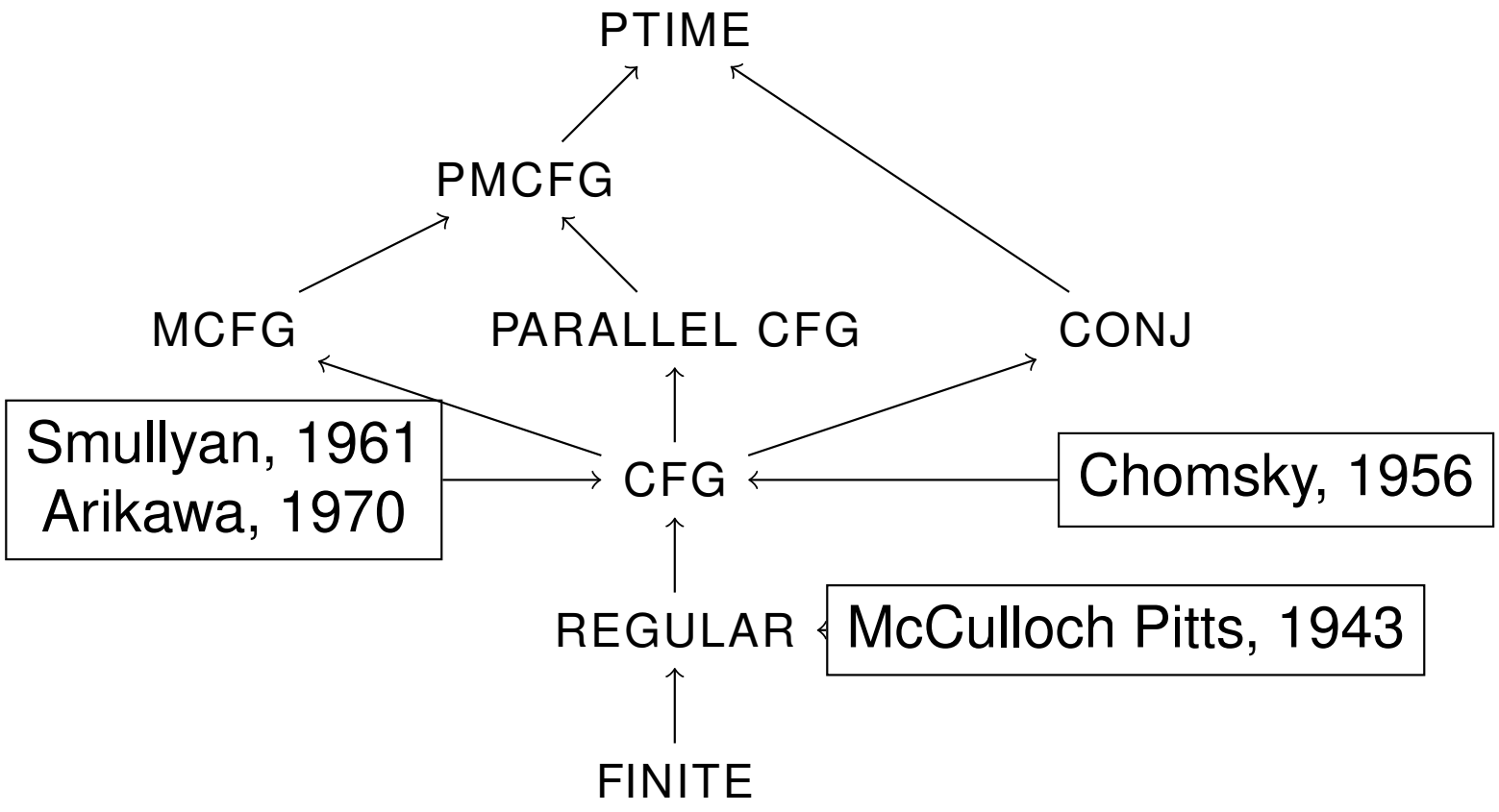
Grammar formalisms



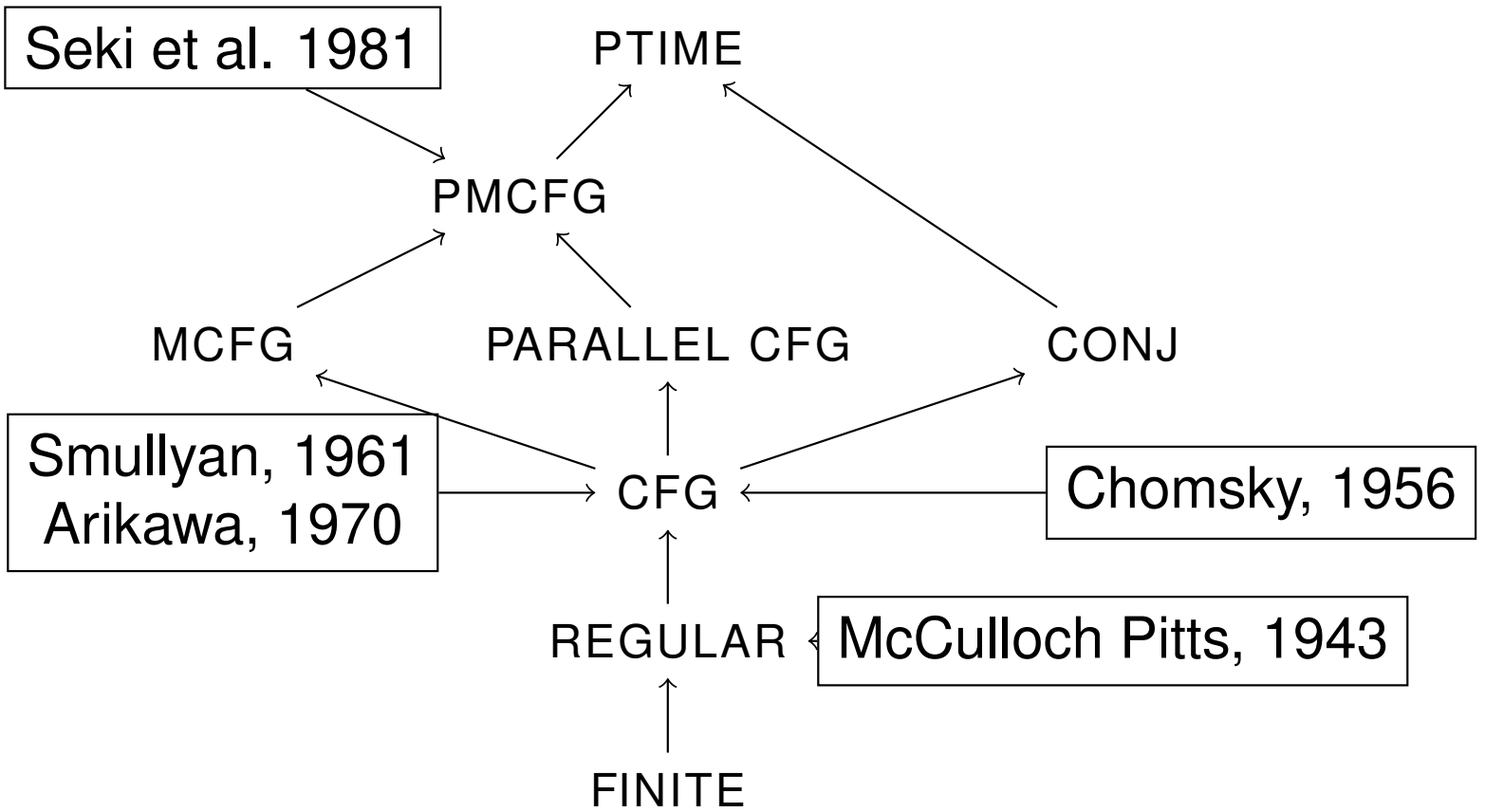
Grammar formalisms



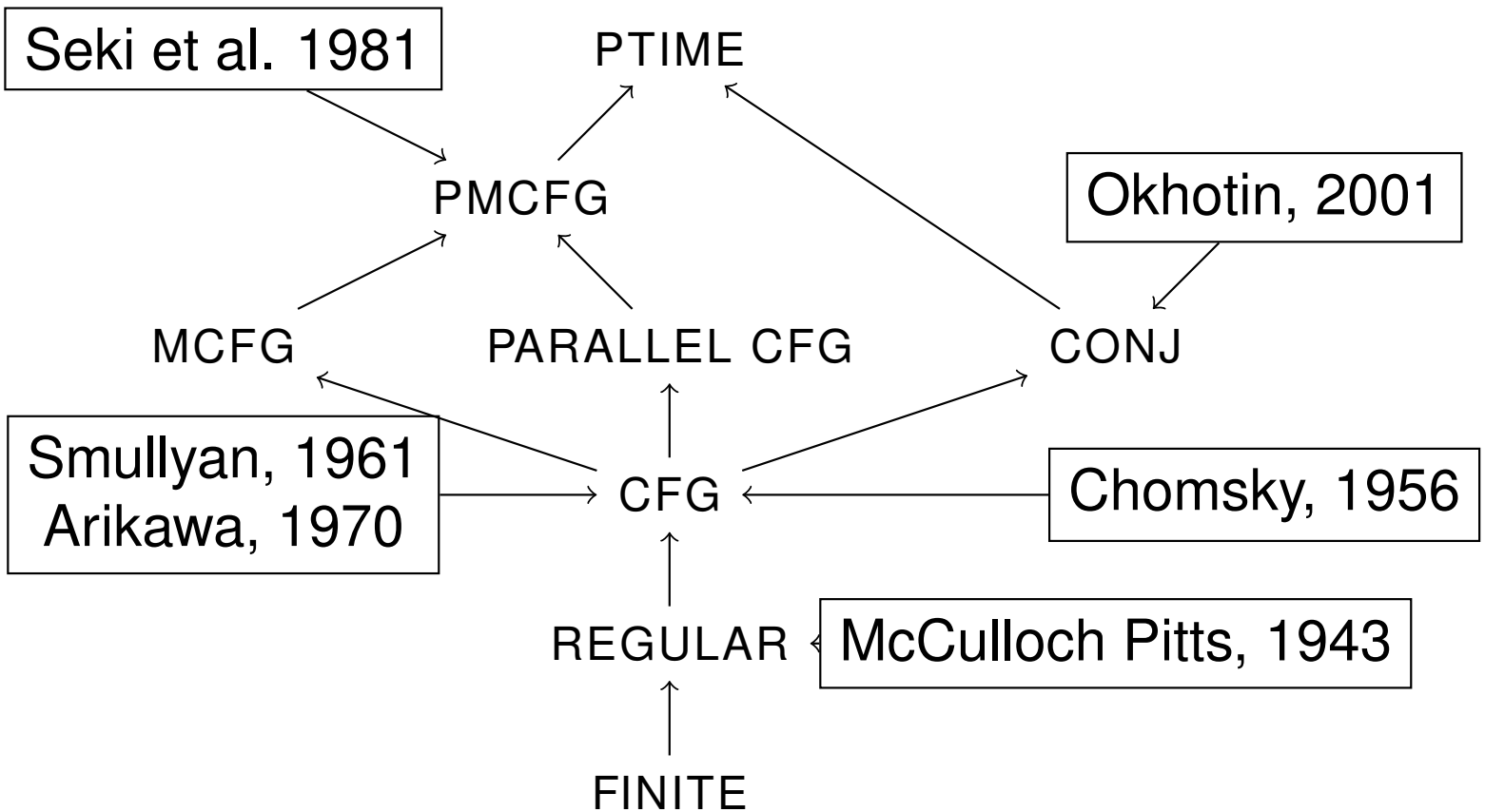
Grammar formalisms



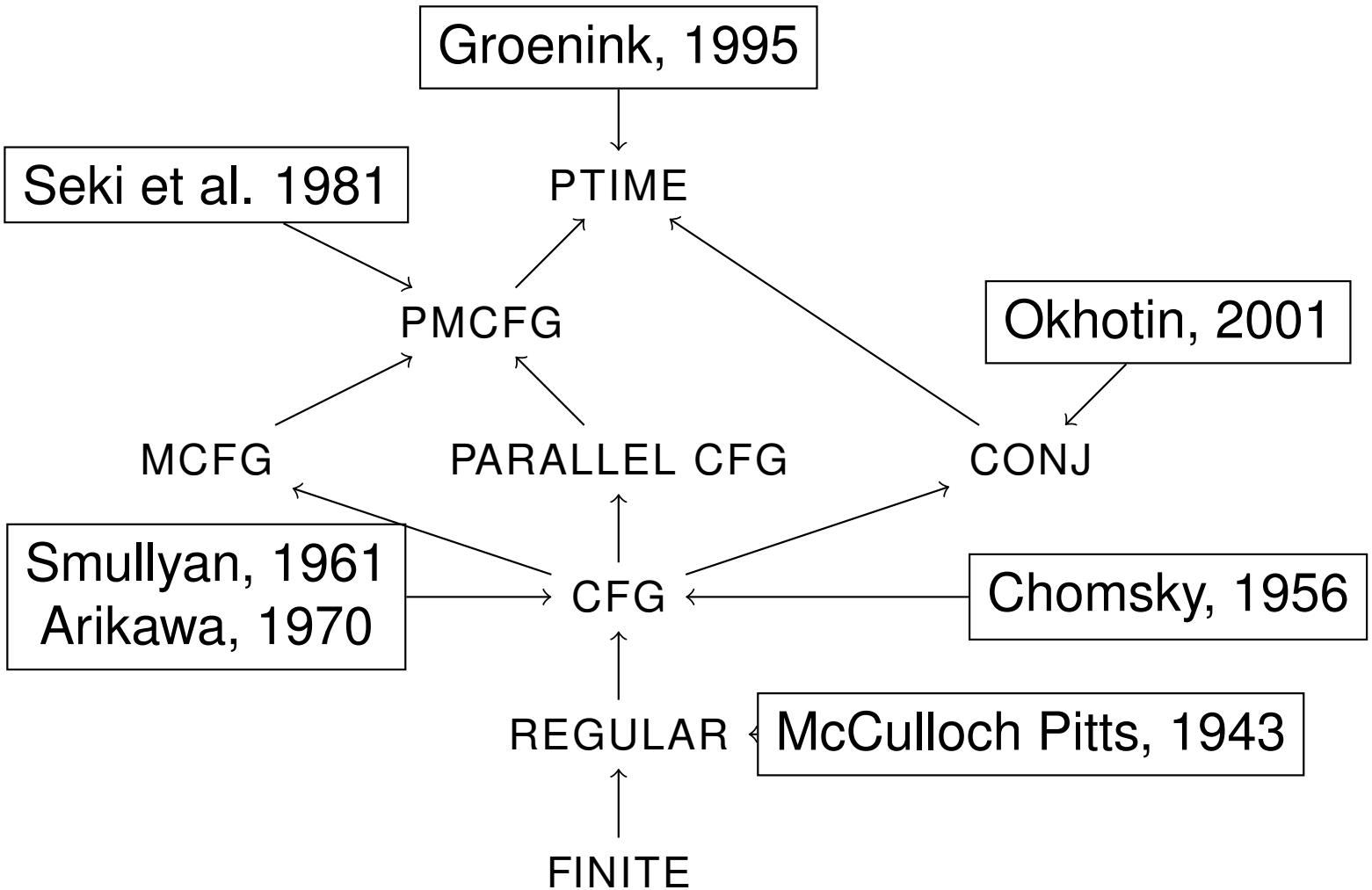
Grammar formalisms



Grammar formalisms

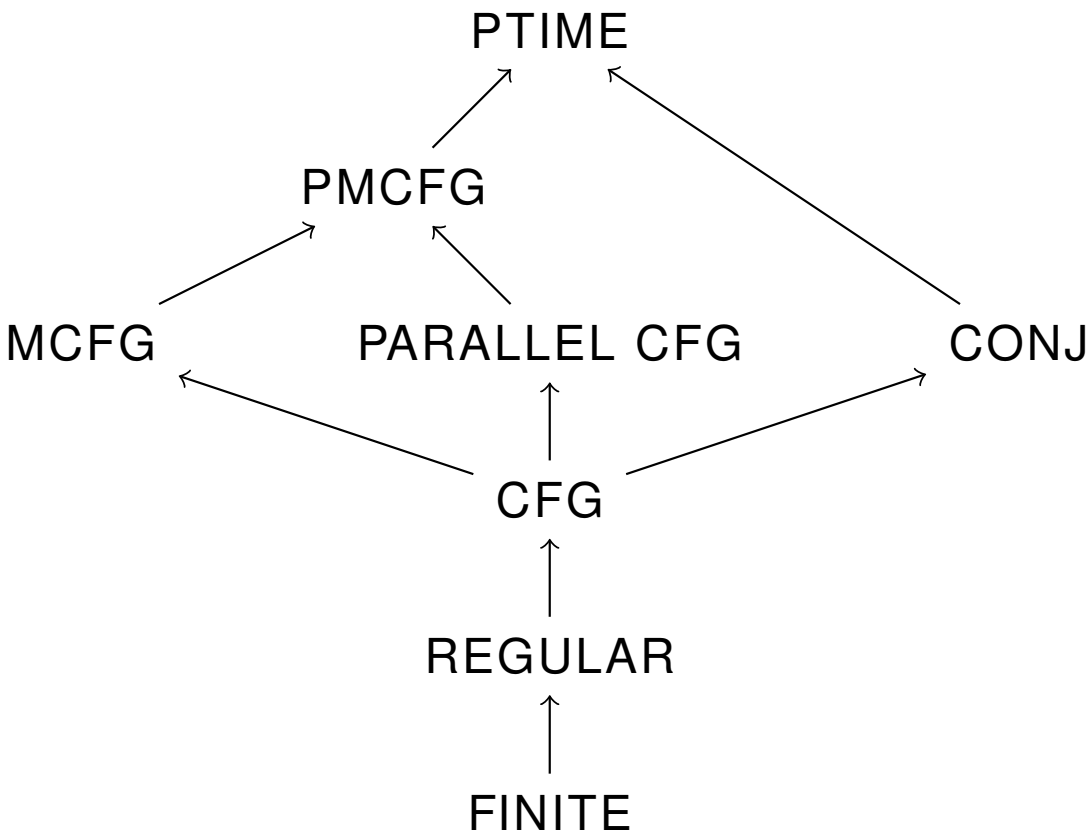


Grammar formalisms



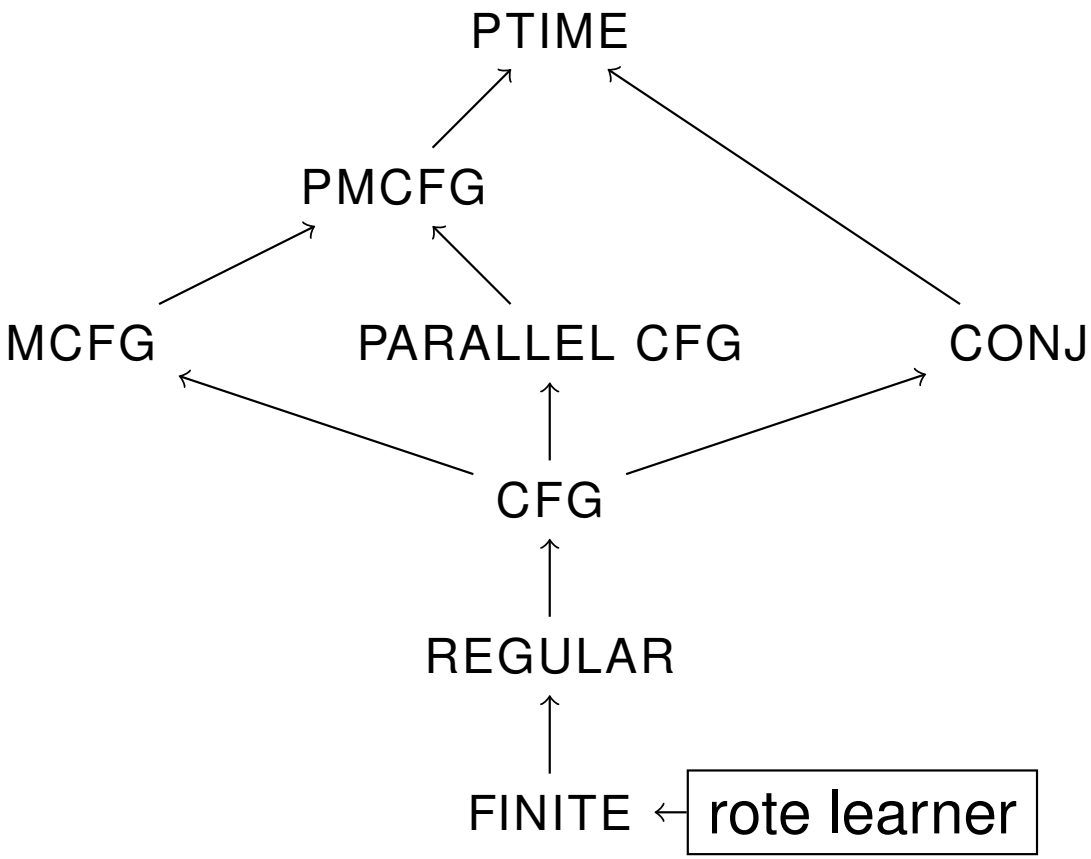
Chomsky hierarchy

Bottom up: learning results



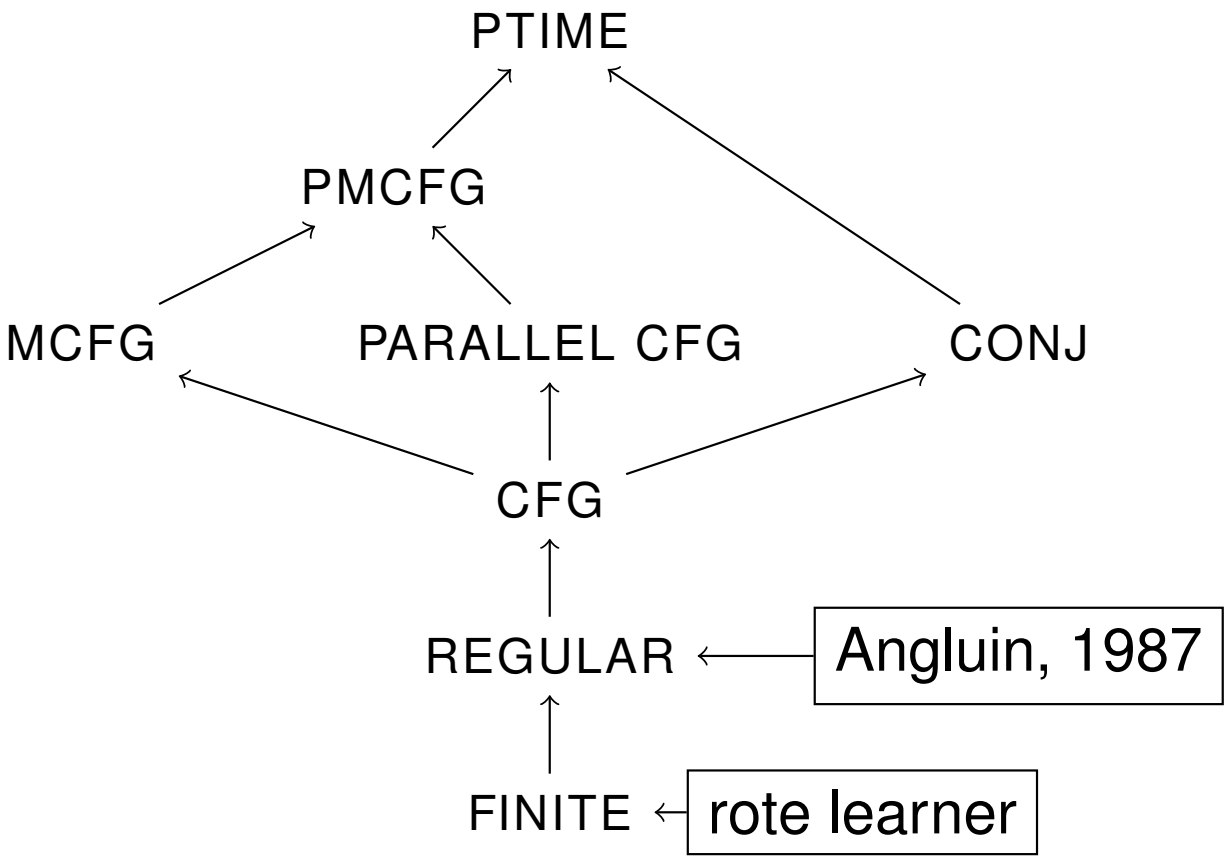
Chomsky hierarchy

Bottom up: learning results



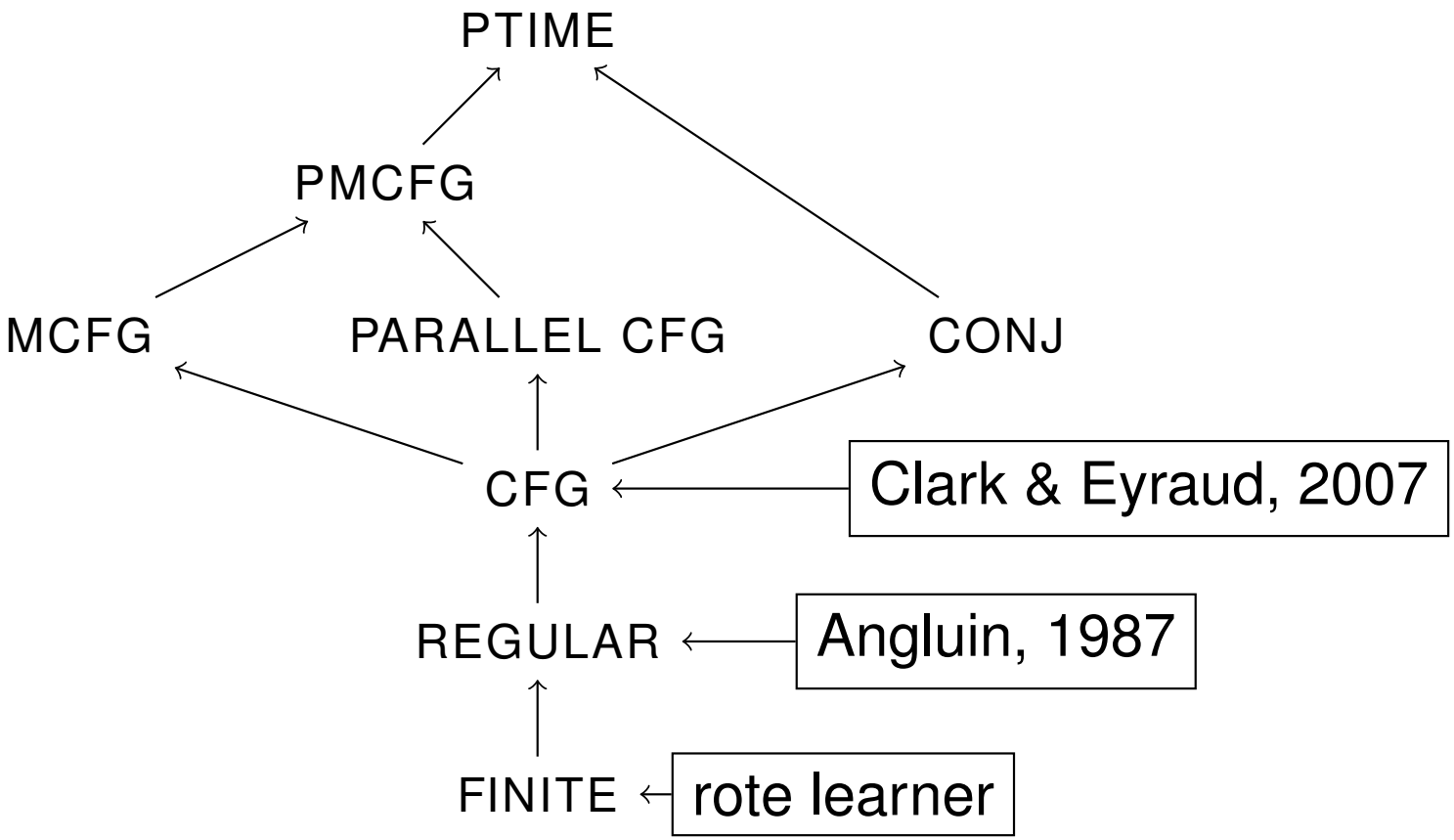
Chomsky hierarchy

Bottom up: learning results



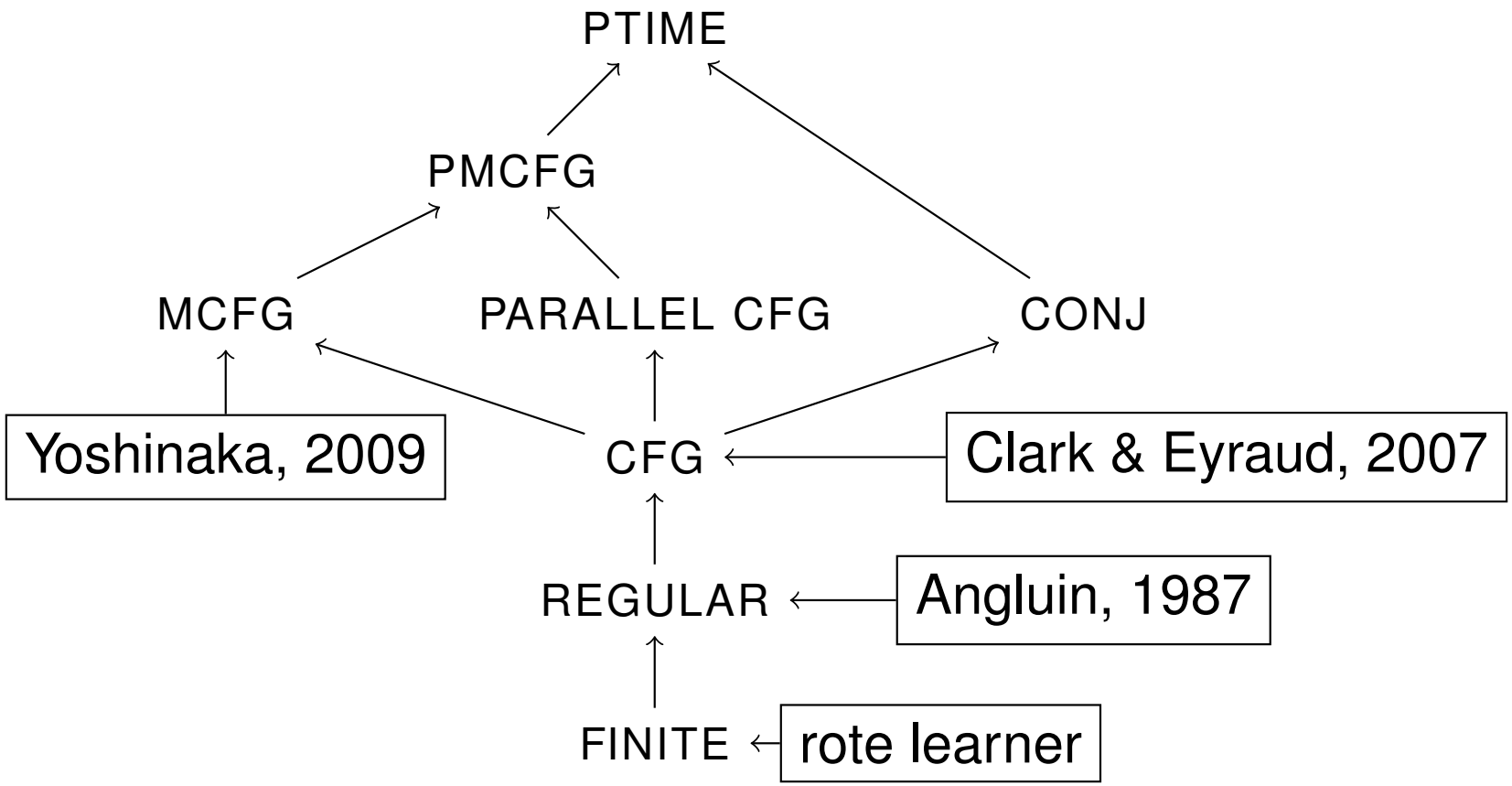
Chomsky hierarchy

Bottom up: learning results



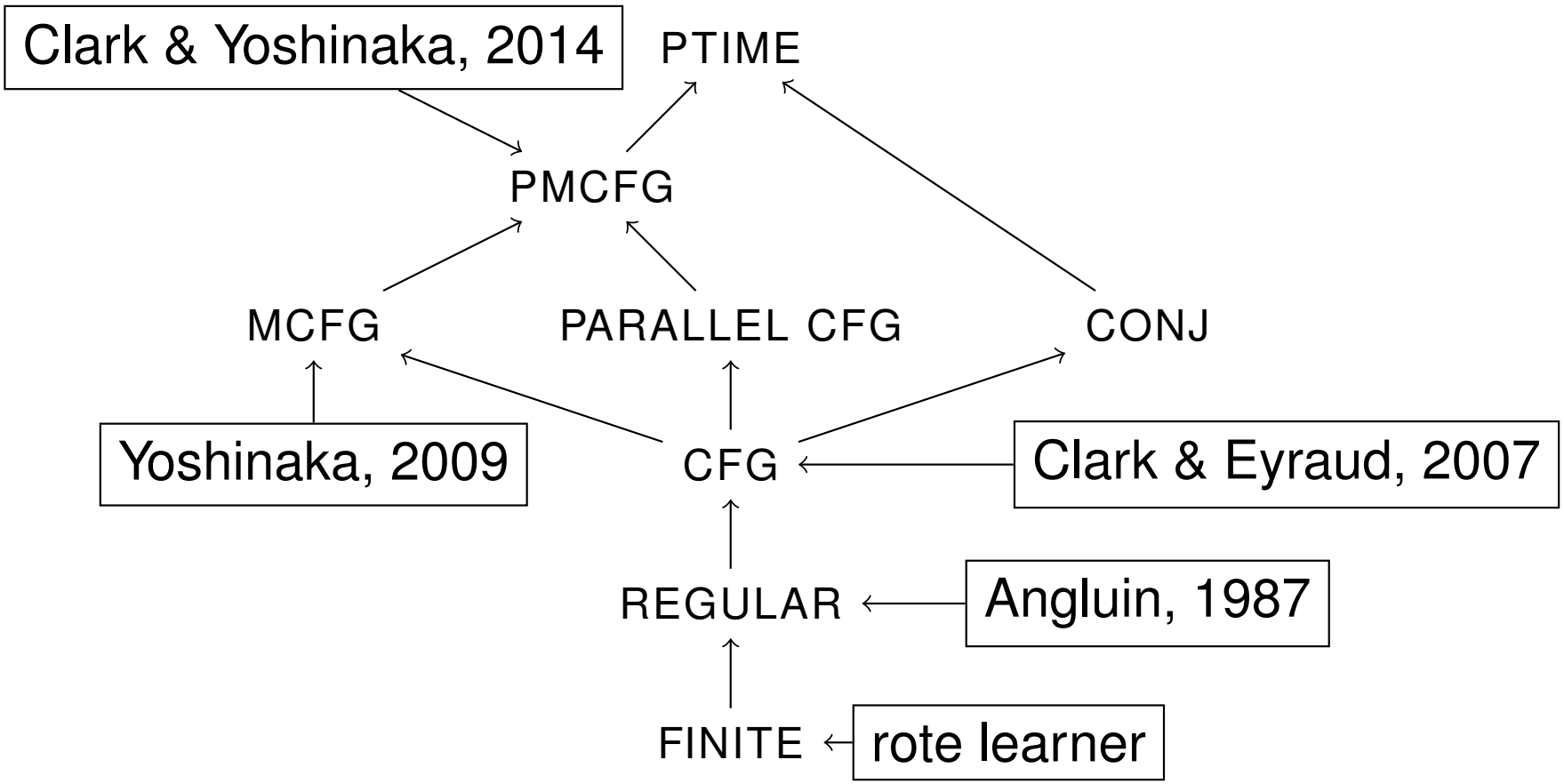
Chomsky hierarchy

Bottom up: learning results



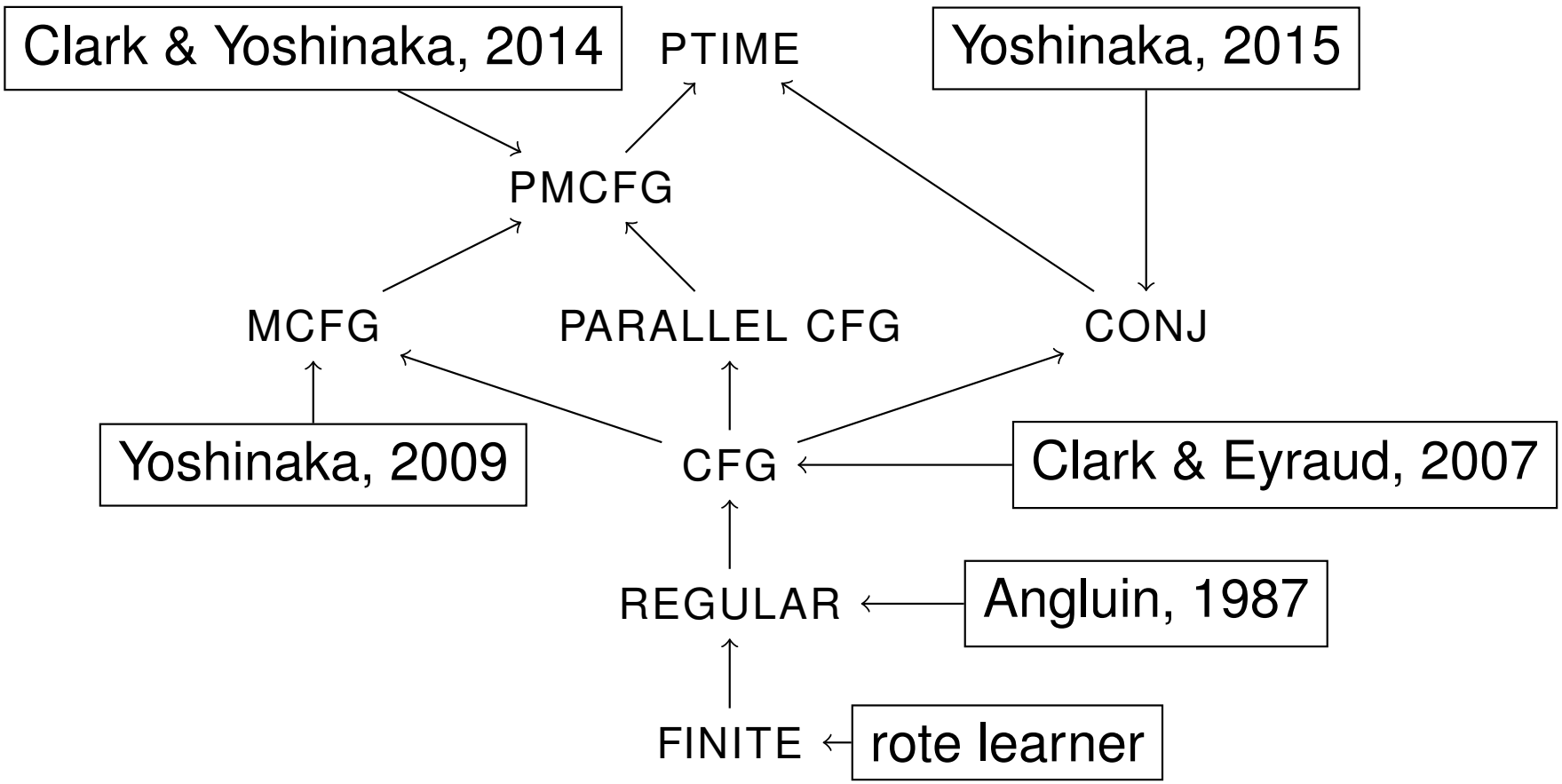
Chomsky hierarchy

Bottom up: learning results



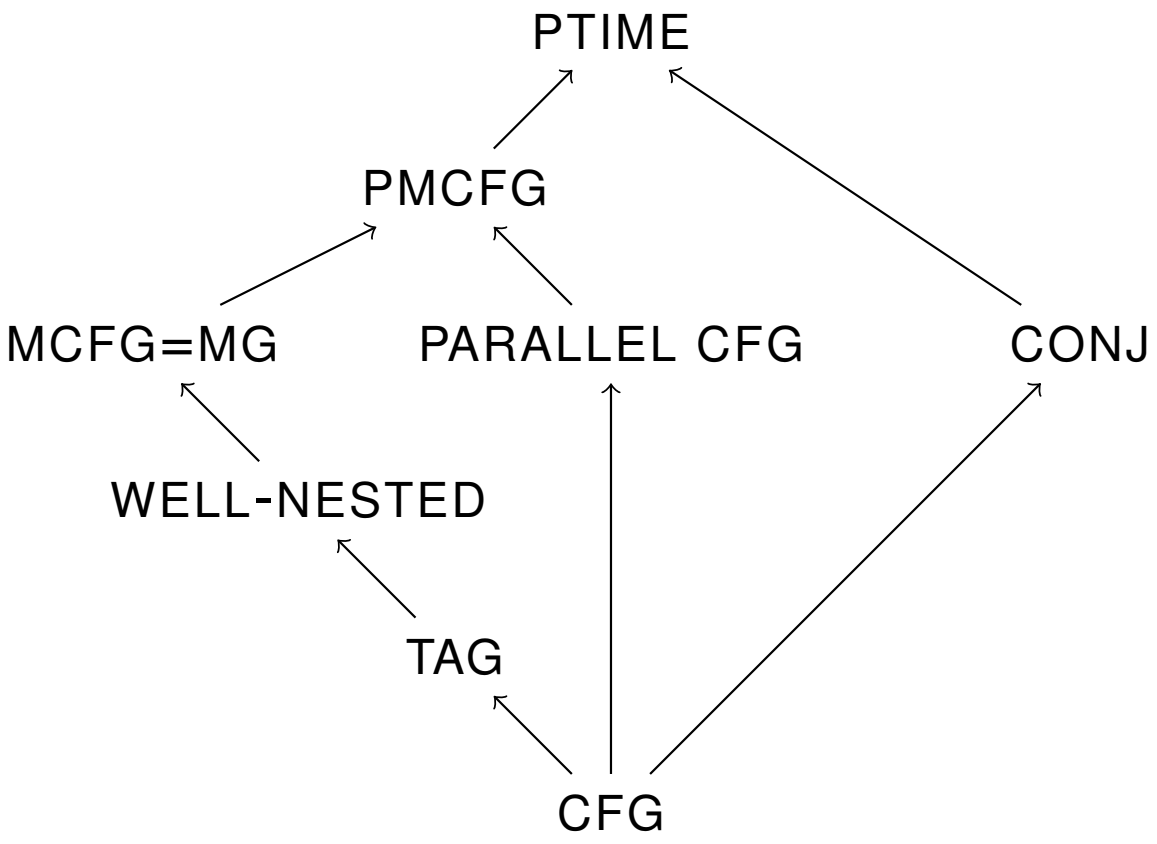
Chomsky hierarchy

Bottom up: learning results



Chomsky hierarchy

Bottom up



Representation

[Stabler, 2013]

This consensus is stable and rather well understood.

- MCFGs are weakly and strongly equivalent to MInimalist Grammars (Stablerian)
- Caveat: the MCFGs may be much bigger than the equivalent MG.

Bottom up notation for CFGs

Top down notation

$$S \rightarrow ab$$

$$S \rightarrow aSb$$

Bottom up

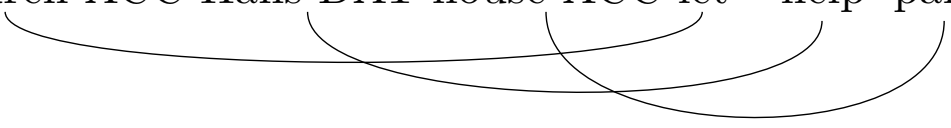
$$S(ab)$$

$$S(axb) \leftarrow S(x)$$

Swiss German

Shieber (1985)

... das mer d'chind em Hans es huus lönd hälfe aastriiche
... that we the children-ACC Hans-DAT house-ACC let help paint



'... that we let the children help Hans paint the house'

Proof by intersection with regular language

$$\{a^n b^m c^n d^m | m, n \geq 0\}$$

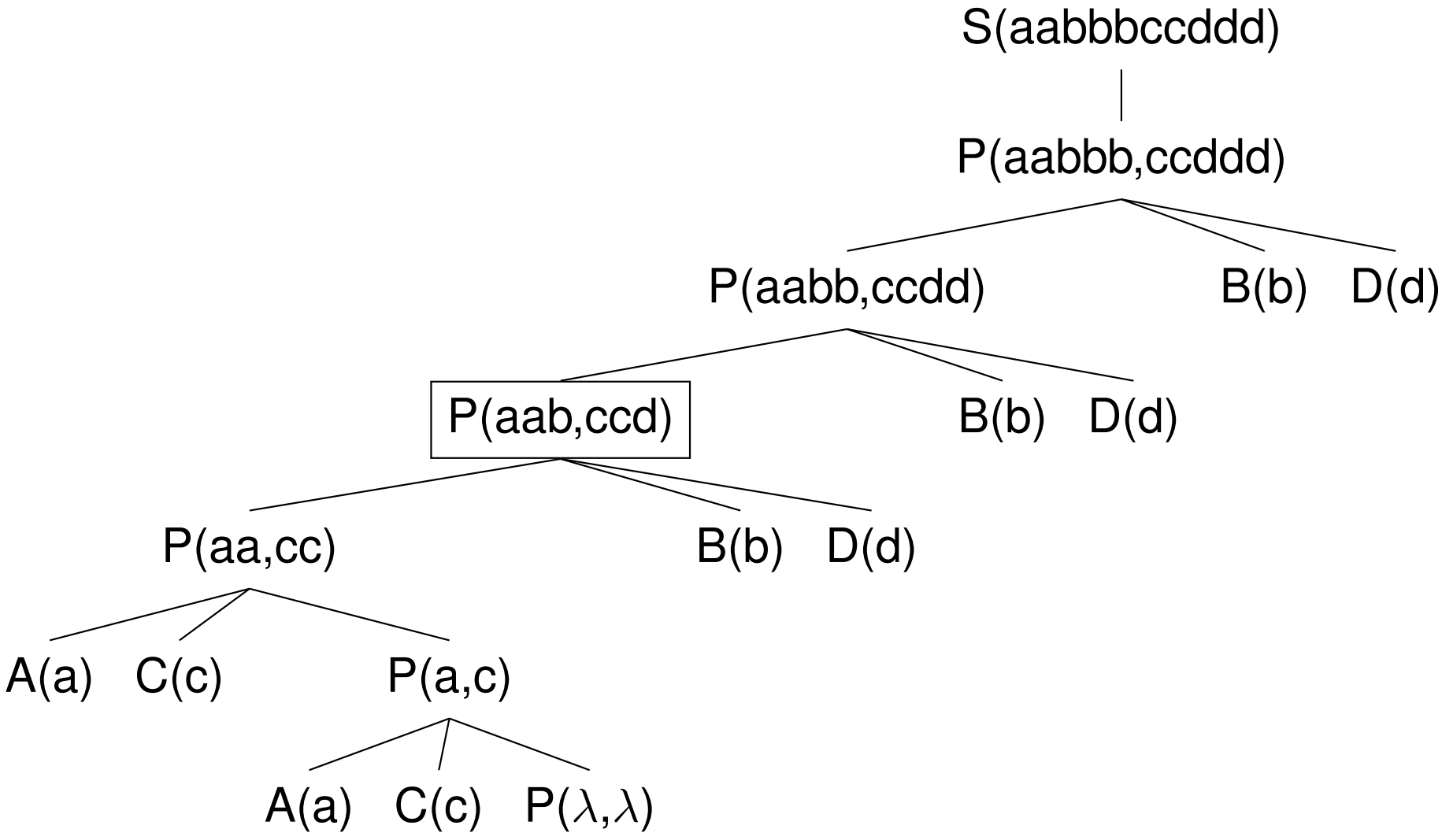
MCFGs

A neutral extension of CFGs.

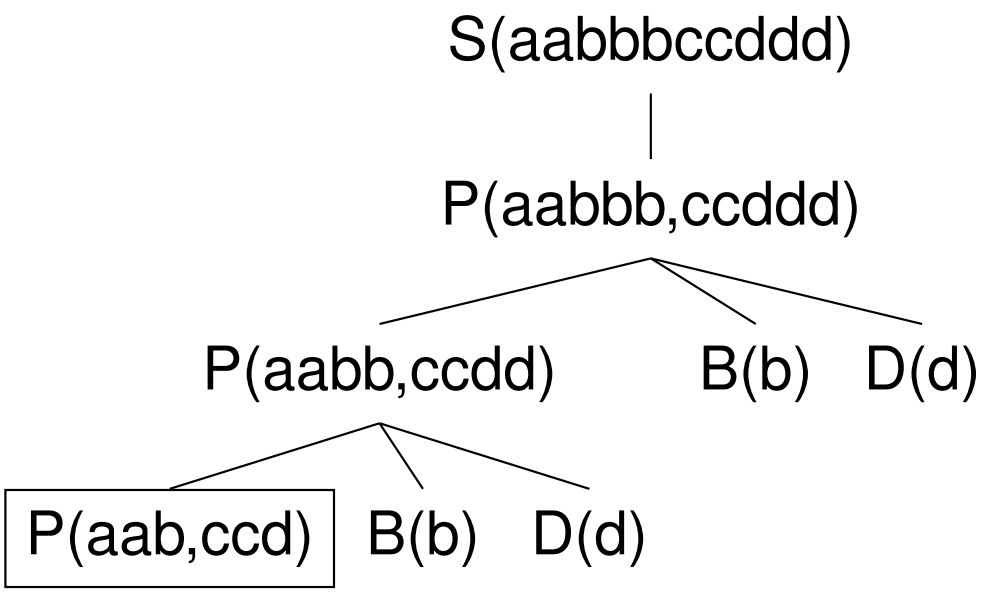
Example grammar

- One nonterminal of dimension 2: P
  $P(\lambda, \lambda)$
  $P(xu, yv) \leftarrow P(x, y)B(u)D(v)$
  $P(ux, vy) \leftarrow A(u)C(v)P(x, y)$
- Lexical rules introducing the terminal symbols:
  $A(a), B(b), C(c), D(d)$
- Start symbols S always has dimension 1.
  $S(xy) \leftarrow P(x, y)$

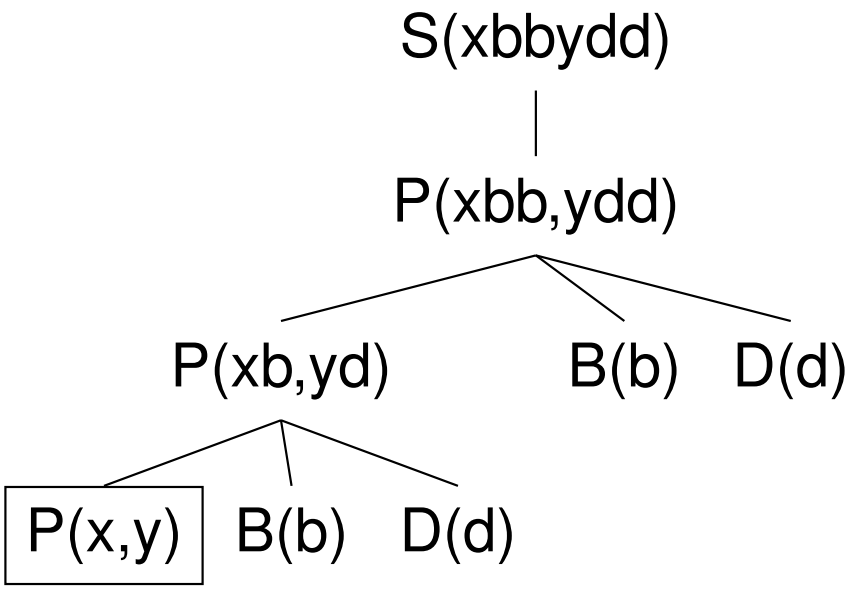
derivation


S(aabbbccddd)
P(aabbb,ccddd)
P(aabb,ccdd)
B(b)
D(d)
P(aab,ccd)
B(b)
D(d)
P(aa,cc)
B(b)
D(d)
A(a)
C(c)
P(a,c)
A(a)
C(c)
P(λ,λ)

derivation



derivation



Derivation contexts of nonterminals of dimension two have two gaps $\square bb \square dd$.

Hierarchy

$\mathbb{G}(p, q)$ where

- p is the maximal dimension
- r is the maximal number of symbols on the right hand side of a rule.

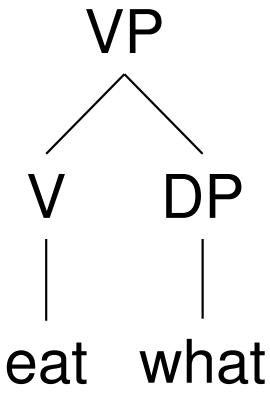
Hierarchy

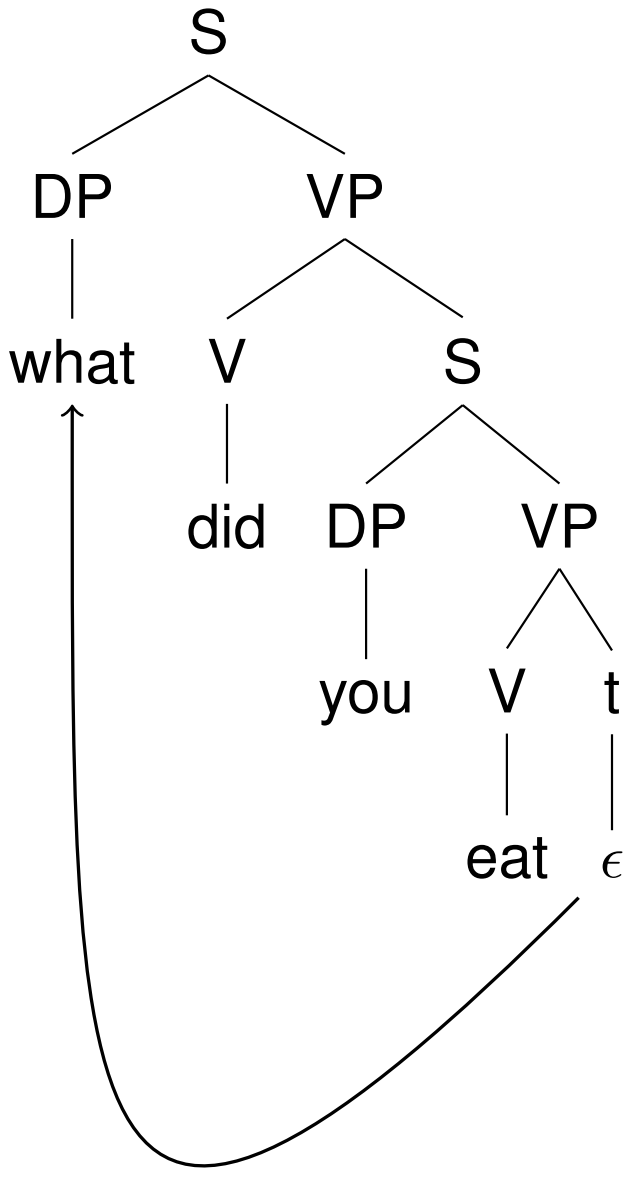
$\mathbb{G}(p, q)$ where

- p is the maximal dimension
- r is the maximal number of symbols on the right hand side of a rule.

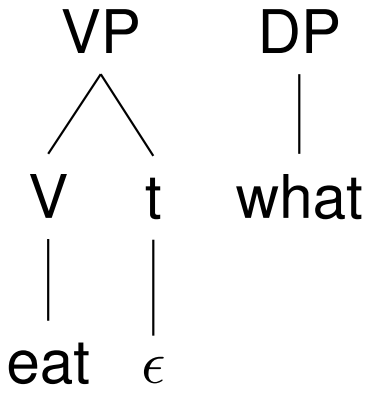
$\mathbb{G}(1, 2)$ is CFGs in Chomsky normal form; $\mathbb{G}(1, *)$ is CFGs.

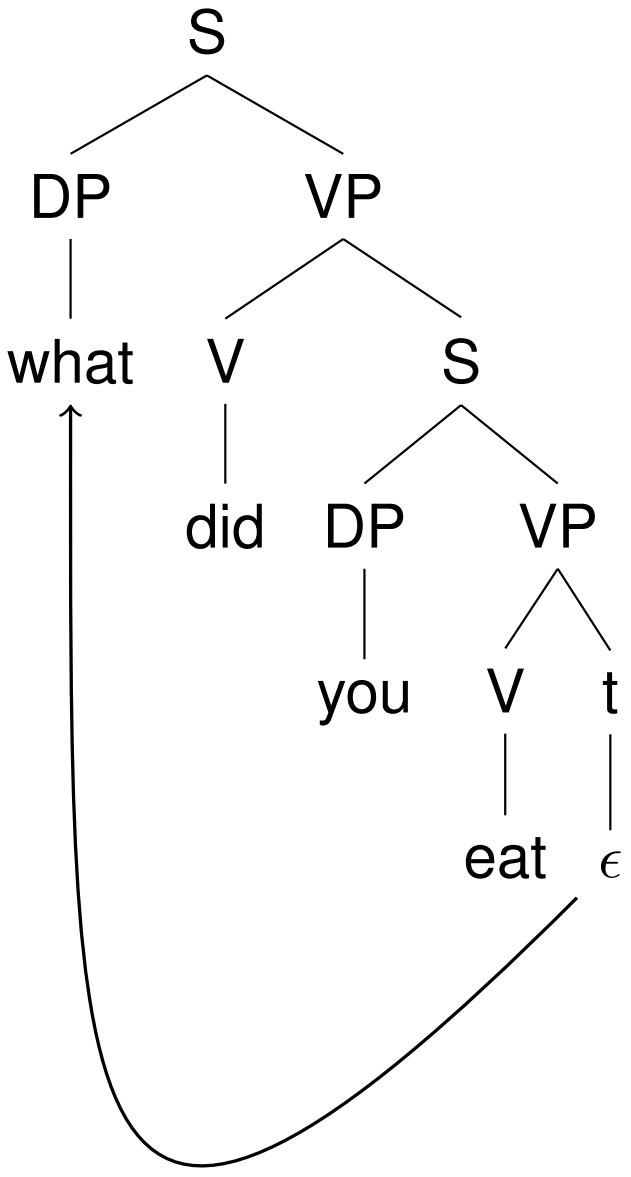
Movement





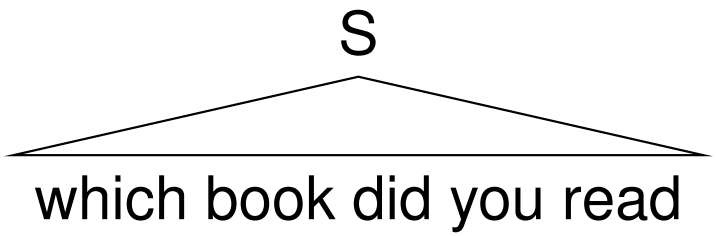
Movement


VP
DP
V
t
what
eat
ϵ


S
DP
VP
what
V
S
did
DP
VP
you
V
t
eat
ϵ

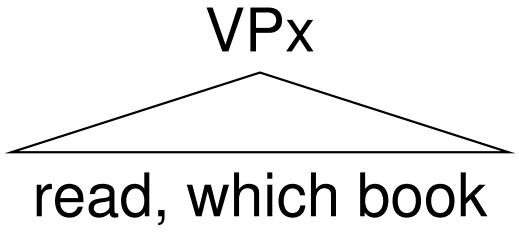
Richer derivations need richer sets of contexts

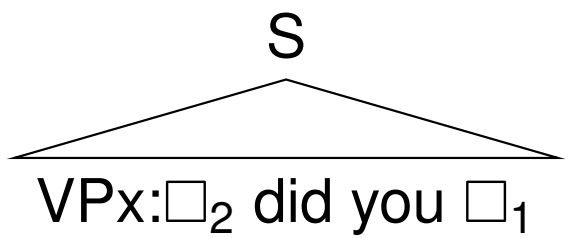
Movement



Richer derivations need richer sets of contexts

Movement





Some nonterminals can generate pairs of strings

Subderivation

Yields a pair of strings ‘eat, what’

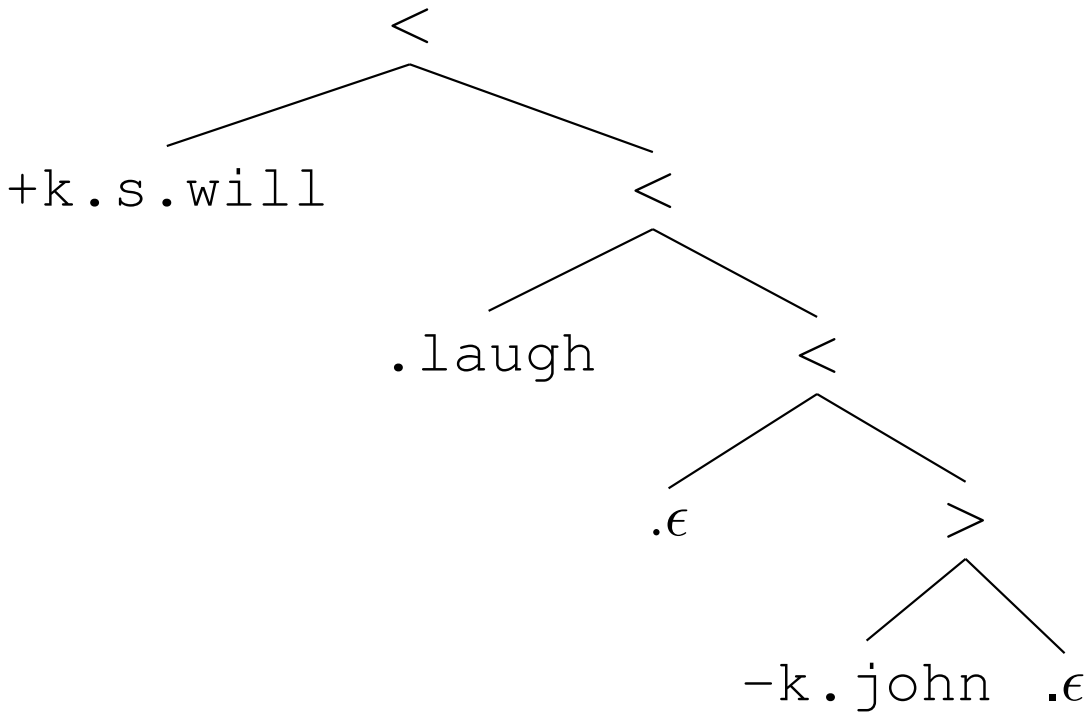
Context

‘eat, what’ $\rightarrow$ ’what did you eat’

$f(x, y) =$ y did you x

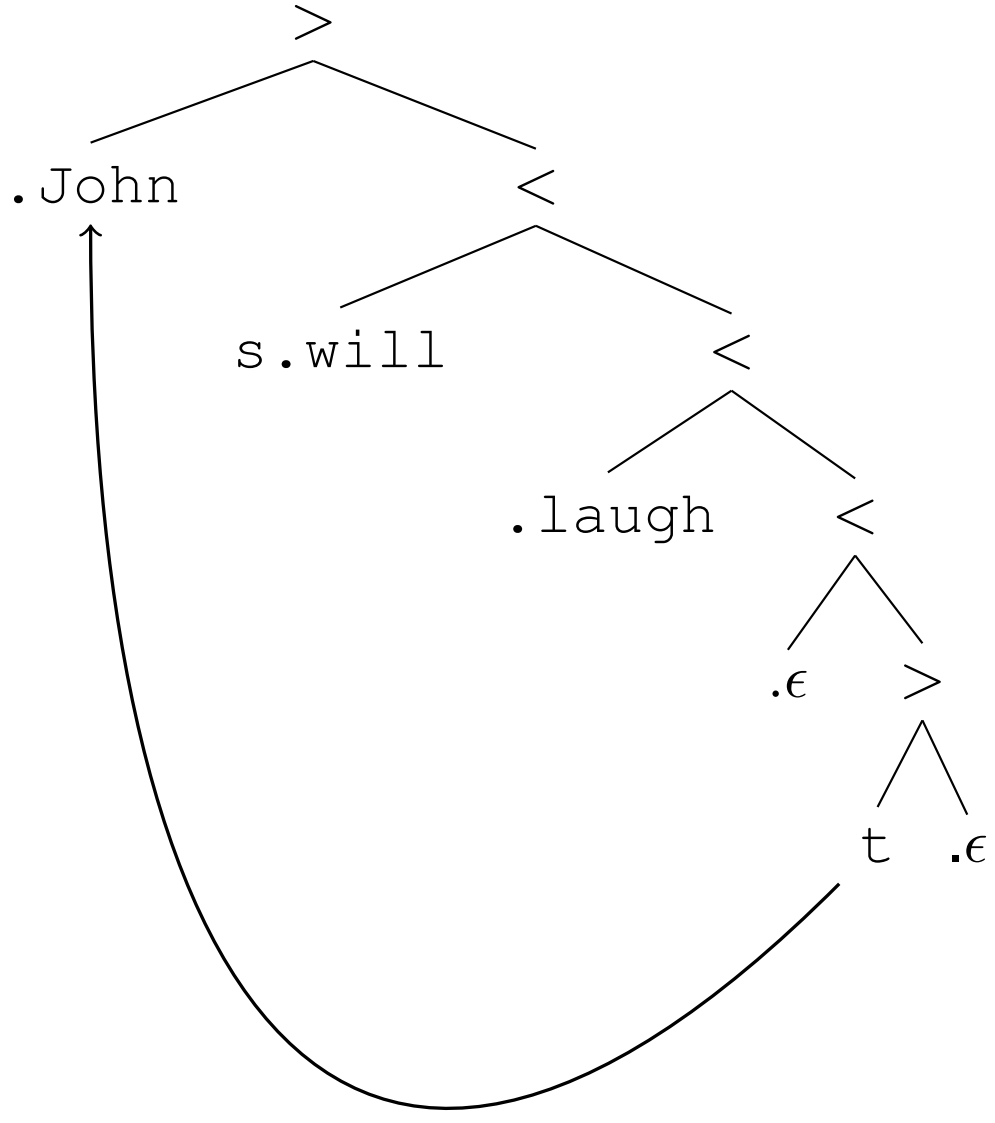
Contemporary Minimalism

short raising to subject



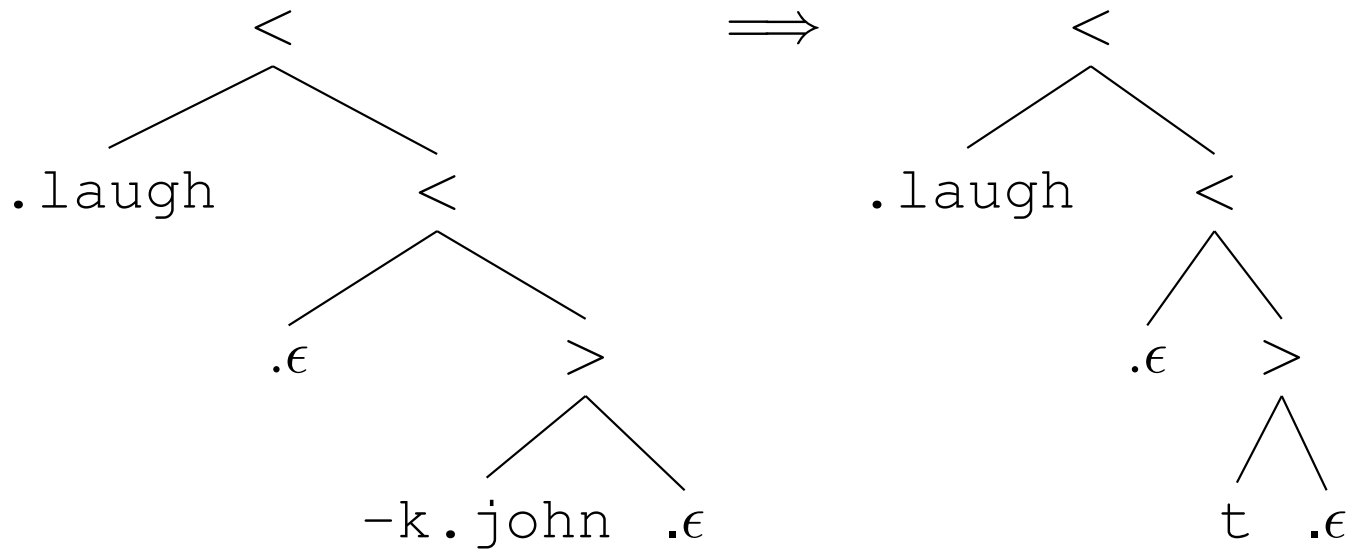
Contemporary Minimalism

short raising to subject



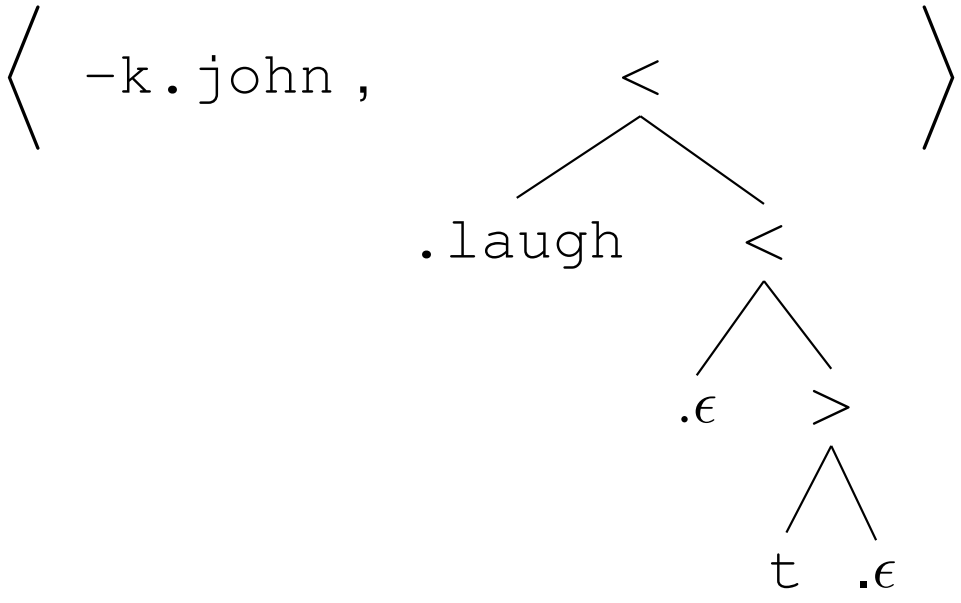
Problem

derived tree changes



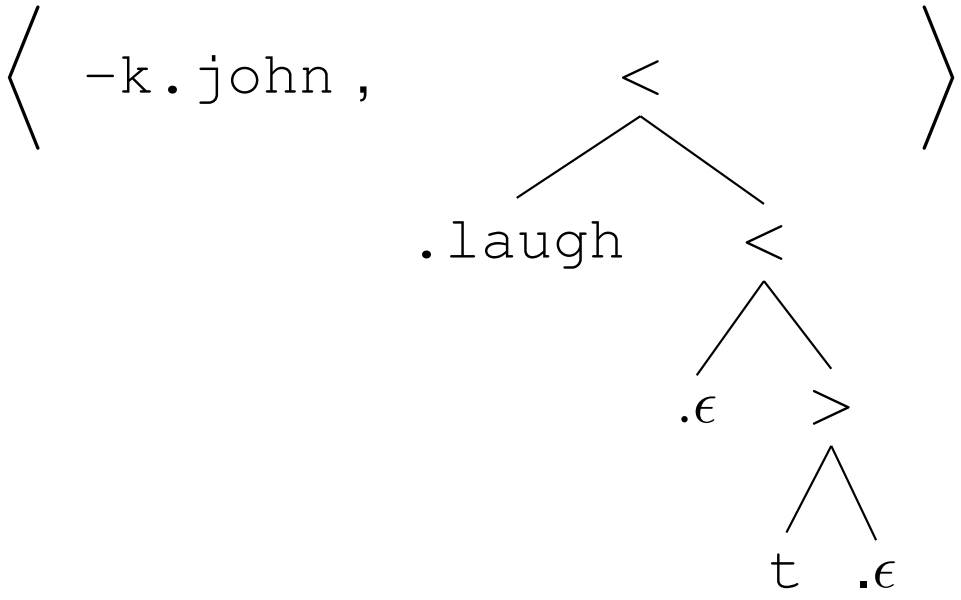
Problem

Make it a tuple



Problem

Make it a tuple

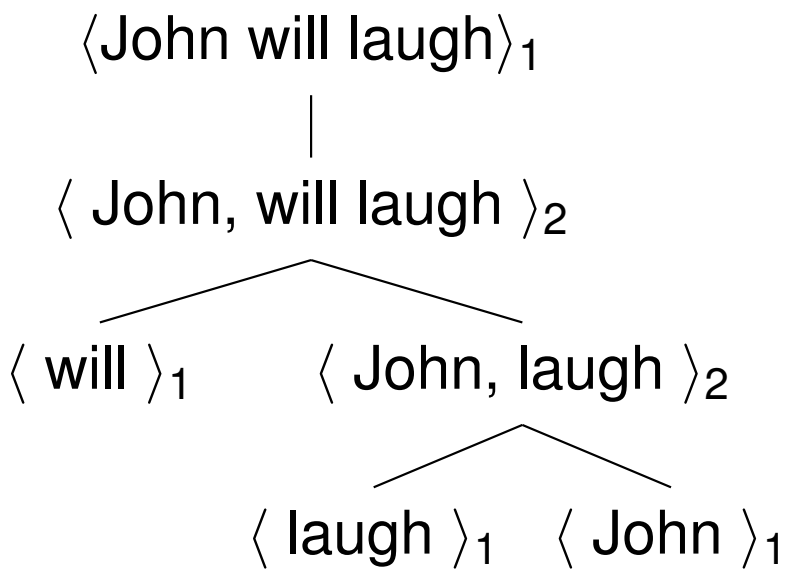


Yields

$\langle \text{john}, \text{laugh} \rangle$

Contemporary Minimalism

short raising to subject



Extending the notion of contexts

Nonterminals of dimension 1

- Yields are strings u
- Contexts have one gap $l\square r$
- Combine them: $l\square r \odot u = lur$

Extending the notion of contexts

Nonterminals of dimension 1

- Yields are strings u
- Contexts have one gap $l\Box r$
- Combine them: $l\Box r \odot u = lur$

Nonterminals of dimension 2

- Yields are pairs of strings $\langle u, v \rangle$
- Contexts have two gaps $l\Box m\Box r$
- Combine them: $l\Box m\Box r \odot \langle u, v \rangle = lumvr$

Finite Context Property

Can we find for each nonterminal a finite set of contexts?

Example: $\{a^n b^m c^n d^m | m, n > 0\}$

P of dimension 2

$$\mathcal{L}(P) = \{\langle a^n b^m, c^n d^m \rangle \mid m, n \geq 0\}$$

Finite Context Property

Can we find for each nonterminal a finite set of contexts?

Example: $\{a^n b^m c^n d^m | m, n > 0\}$

P of dimension 2

$$\mathcal{L}(P) = \{\langle a^n b^m, c^n d^m \rangle \mid m, n \geq 0\}$$

- $\{a\Box bc\Box d\}^{\triangleleft} = \mathcal{L}(P)$
- $\{a\Box bccd\}^{\triangleleft} = \{a\}$
- $\{\Box\}^{\triangleleft} = L$

Extending the learning results

All of the results for CFGs transfer directly to MCFGs.

Result

We can learn all $\mathbb{G}(p, q)$ with the k-context property efficiently from positive data and MQs.

Semilinearity

Informally

Semilinearity (Joshi, 1991):

> *is intended to be an approximate characterization of the linguistic intuition that sentences of a natural language are built from a finite set of clauses of bounded structures using certain simple linear operations.*

Standard view: natural languages are semilinear.

Semilinearity

More formally

A language L is semilinear iff it is letter equivalent to a regular language.

Semilinear languages

All regular, context-free and multiple context-free languages are semilinear.

Non-semilinear languages

$\{ a^{2^n} \mid n > 0 \}$
$\{ a^{n^2} \mid n > 0 \}$

Copying

Kobele, 2006

Copying exists. There are constructions in natural language that require reference to identity of subparts of expressions for their description. This much, at least, is uncontroversial. What is controversial is the proper locus of explanation of these facts; whether copying should be considered syntactic, phonological, semantic, or extra-grammatical.

Copying

Some controversy over whether copying is needed in natural language:

- Chinese number names
- Case-stacking (*Suffixaufnahme*) in Australian language and Old Georgian
- Relative clauses in Yoruba and Wolof
- Reduplication in morphology

Copying in the grammar

$$N(xyx) \leftarrow P(x), Q(y)$$

x occurs twice on the left hand side of the grammar.

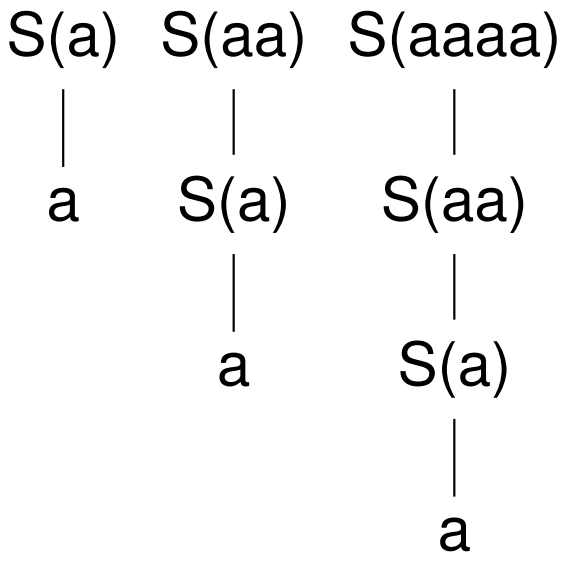
- $P(abc)$
- $Q(e)$
- $\Rightarrow N(abceabc)$

Parallel Multiple Context-Free Grammars

Example grammar

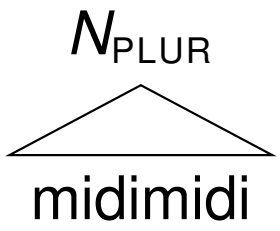
$$S(a)$$

$$S(xx) \leftarrow S(x)$$



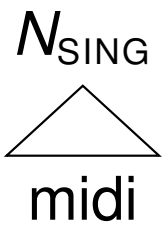
Richer derivations need richer sets of contexts

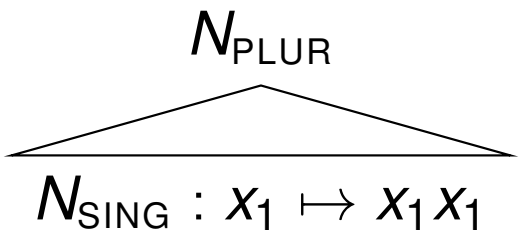
Reduplication in Dyirbal morphology



Richer derivations need richer sets of contexts

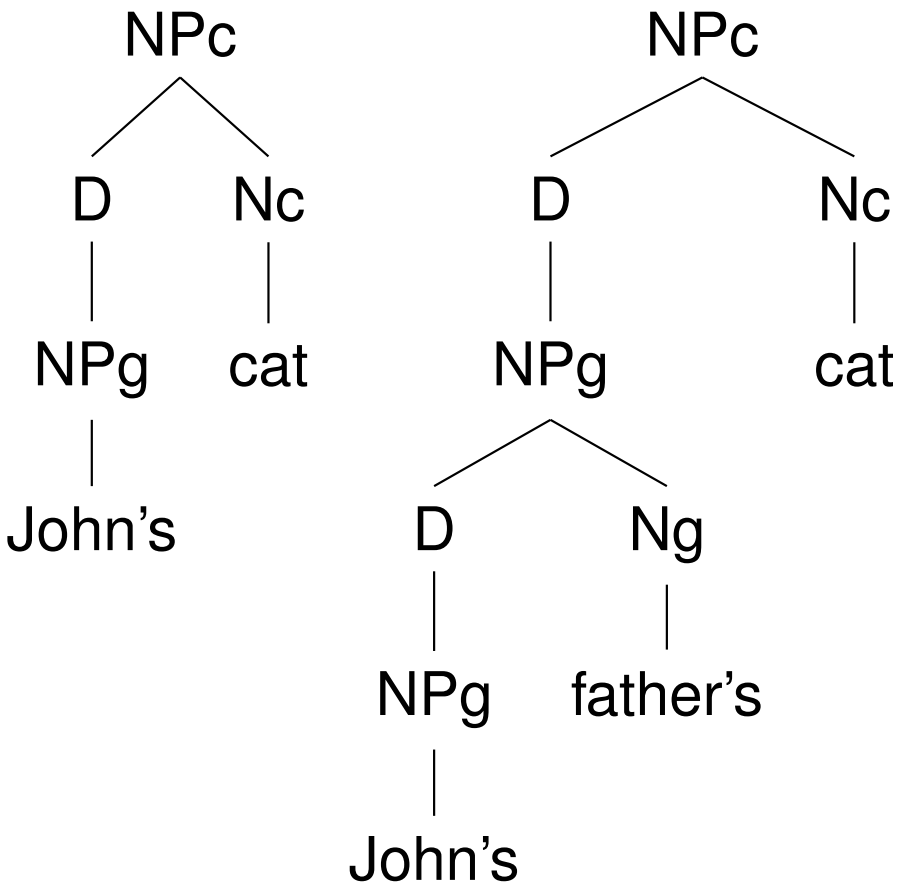
Reduplication in Dyirbal morphology





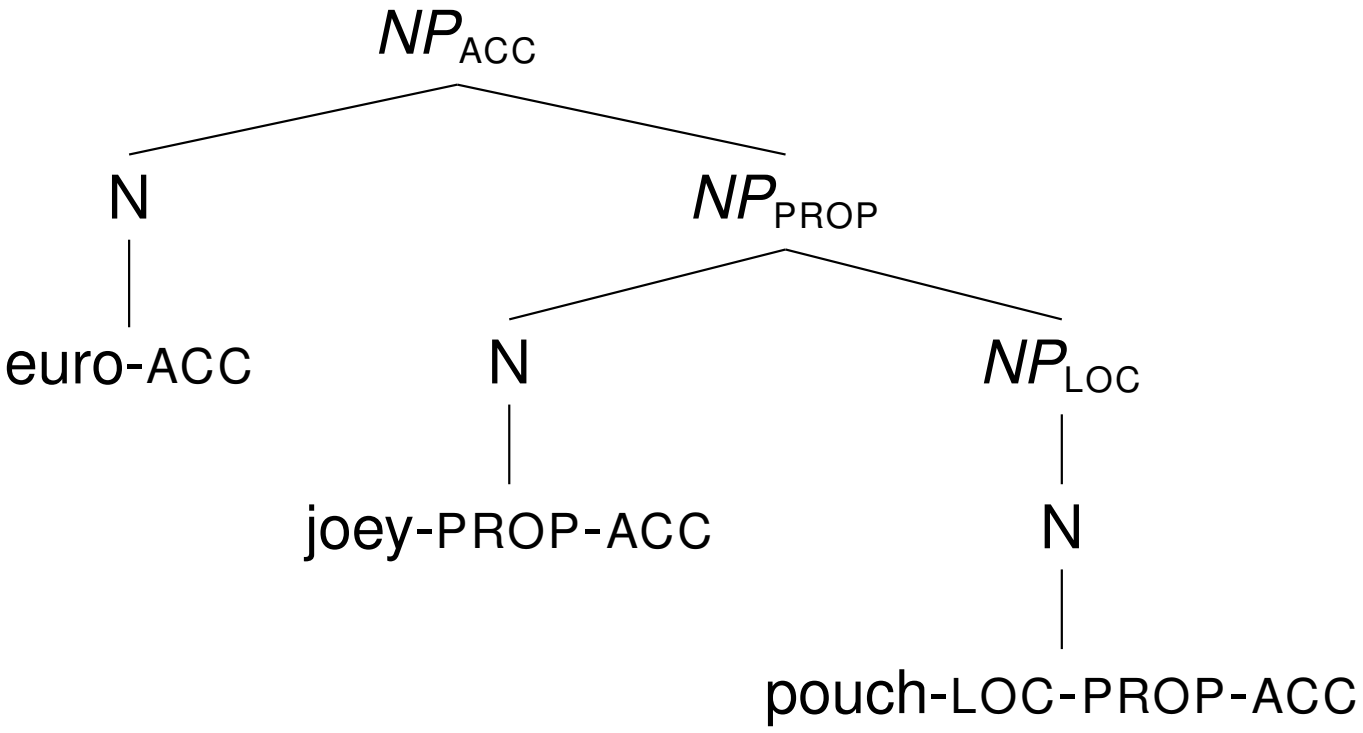
Suffixaufnahme/Case stacking

- John's cat
- John's father's cat
- John's friend's father's cat


NPc
D
Nc
NPg
cat
John’s
NPc
D
Nc
NPg
cat
D
Ng
NPg
father’s
John’s

Case stacking in Martuthunira

Sadler and Nordlinger, 2006



- tharnta-a mirtily-marta-a thara-ngka-marta-a
- euro-ACC joey-PROP-ACC pouch-LOC-PROP-ACC
- “I saw the euro with a joey in its pouch.”

Suffixaufnahme/Case stacking

If English had case-stacking . . .

- John's cat
- John's's father's cat
- John's's's friend's's father's cat
- John's's's's friend's's's father's's cat's tail

Suffixaufnahme/Case stacking

If English had case-stacking . . .

- John's cat 1
- John's's father's cat 3
- John's's's friend's's father's cat 6
- John's's's's friend's's's father's's cat's tail 10

Suffixaufnahme in Old Georgian

Michaelis and Kracht

Alphabet

$\{n, g, v\}$

Language

$\{nv, nngv, nngnggv, nngnggngggv, \ldots\}$.

- $nn \underbrace{g} n \underbrace{gg} n \underbrace{ggg} v$

Yoruba relative clauses

Kobele, 2006

Recursive copying in Yoruba

'rira NP ti Ade ra NP ko da'
The fact that Ade bought NP is not good.

- NP must be copied
- It can contain relative clauses that must also be copied.

Trivial example of a non-semilinear language

$L = \{a^{2^n} \mid n > 0\}$

$S(x_1 x_1) \leftarrow S(x_1)$

$S(a) \leftarrow$

$S(aaaaaaaa)$

|

$S(aaaa)$

|

$S(aa)$

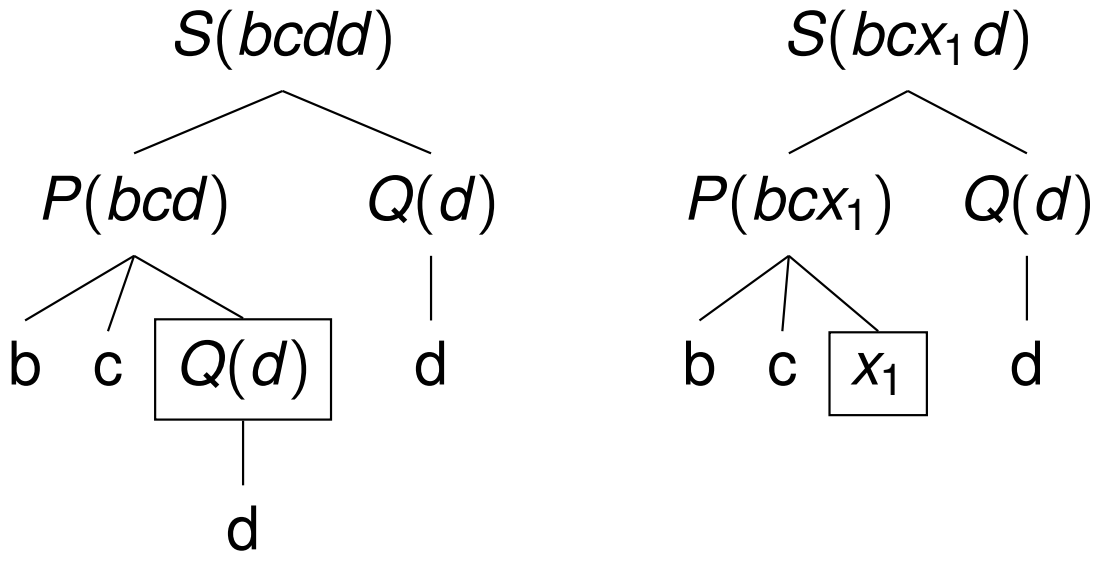
|

$S(a)$

$$S(nngngggv)$$
$$|$$
$$N(nngn, gg)$$
$$|$$
$$N(nn, g)$$
$$|$$
$$N(n, \lambda)$$

$$S(x_1 x_2 v) \leftarrow N(x_1, x_2)$$
$$N(x_1 x_2 n, x_2 g) \leftarrow N(x_1, x_2)$$
$$N(n, \lambda) \leftarrow \ .$$

Context in a CFG derivation

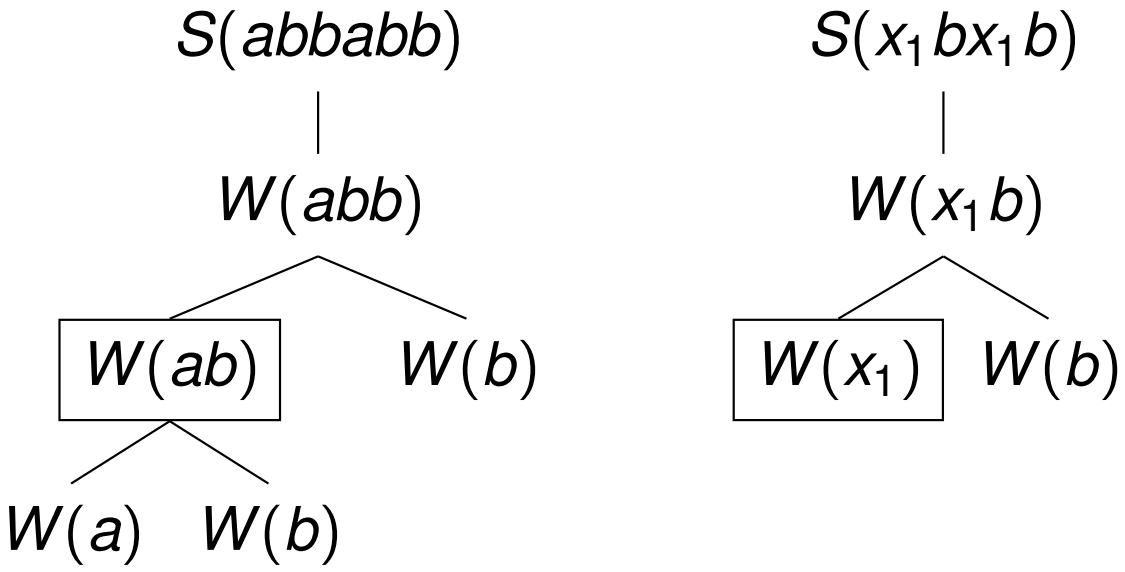


Function

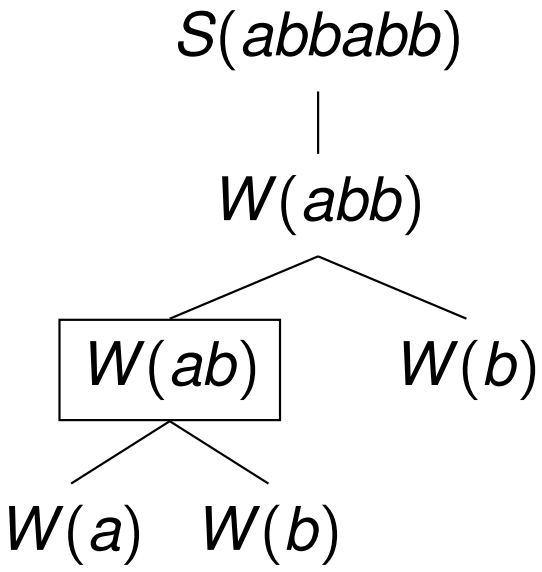
$f(x_1) = bcx_1d$

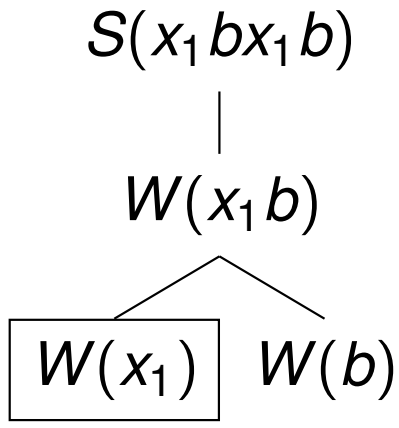
The context (bc, d) or $bc\square d$

Generalised Context



Generalised Context





- $f(x_1) = x_1 b x_1 b$
- $f(ab) = abbabb$

Contexts

1-context: CFG

$lx_1 r$

2-context:MCFG

$lx_1 mx_2 r$

2-copying 2-context:PMCFG

$lx_1 mx_1 x_2 r$ or $x_1 x_1 px_2 x_2$

We need r-copying d-contexts for nonterminals of dimension d.

Generalised distribution

- We say that u occurs in the context lx_1x_1r if $luur \in L_*$

Generalised distribution

- We say that u occurs in the context lx_1x_1r if $luur \in L_*$
- We say that the d-word $\vec{v}$ occurs in the r-copying d-context c if $c[\vec{v}] \in L_*$.

Example

Copy language

$W(a) \leftarrow ;$
$W(b) \leftarrow ;$
$W(x_1 x_2) \leftarrow W(x_1), W(x_2).$
$S(x_1 x_1) \leftarrow W(x_1)$

- Consider the 2-copying 1-context $x_1 x_1$
- v occurs in $x_1 x_1$ iff $\vdash_G W(v)$.
- $C_W = \{x_1 x_1\}$
- $\mathcal{L}(G, W) = \{ v \in \Sigma^* \mid C_W[v] \subseteq \mathcal{L}(G) \}$

We can find a context that picks out exactly $\mathcal{L}(G, W)$

Learnability constraint

We say that a PMCFG G has *the* (r, s)*-finite context property* ((r, s)-FCP) if each nonterminal $A \in N_d$ admits a nonempty set C_A of r-copying d-contexts such that $|C_A| \leq s$ and

$$\mathcal{L}(G, A) = \{ \vec{v} \mid C_A[\vec{v}] \subseteq \mathcal{L}(G) \}.$$

Old Georgian

Context $f_N(x_1, x_2) = x_1 x_2 n x_2 g v$

- $f_N(x_1, x_2) = nngnggngggv$ implies $x_1 = nngn$, $x_2 = gg$.
- $f_N(x_1, x_2) = nngnggv$ implies $x_1 = nn$, $x_2 = g$.
- $f_N(x_1, x_2) = nngv$ implies $x_1 = n$, $x_2 = \lambda$.

We derive $N(x_1, x_2)$ iff $f_N(x_1, x_2) \in L$.

$$C_A = \{x_1 x_2 n x_2 g v\}$$

$$\mathcal{L}(G, A) = \{ \vec{v} \mid C_A[\vec{v}] \subseteq \mathcal{L}(G) \}$$

So this has the $(2, 1)$-FCP.

PMCFGs

Parallel Multiple Context-Free Grammars

Clark and Yoshinaka (2013) MLJ

A hierarchy of PMCFGs can be identified in the limit using positive data and membership queries.
$\mathbb{G}(p, q, r, s)$.

- This seems to contain all natural languages.
- More precisely: There are no arguments that the class of natural languages does not lie in one of these classes.
- Problems:
  - Only dual learning! combinatorial problems with primal approach.
  - Weak learning only
  - Uses membership queries
  - No notion of “feature”

Discussion

- Is this enough?
- What does learnability mean here?

Bibliography I


Stabler, E. P. (2013).
The epicenter of linguistic behavior.
In Sanz, M., Laka, I., and Tanenhaus, M. K., editors, *Language Down the Garden Path: The Cognitive and Biological Basis of Linguistic Structures*, pages 316–323. Oxford University Press.