Beyond Context-Free grammars

Theoretical Linguistics and Formal Learning Theory

Alexander Clark

Manchester

6 November 2013

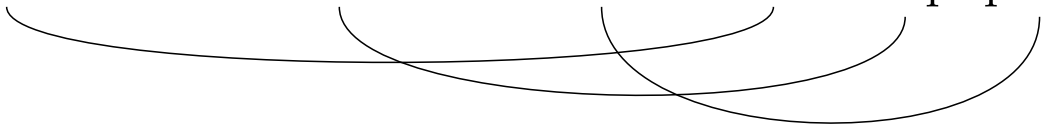
Outline

Outline

Swiss German

Shieber, 1985

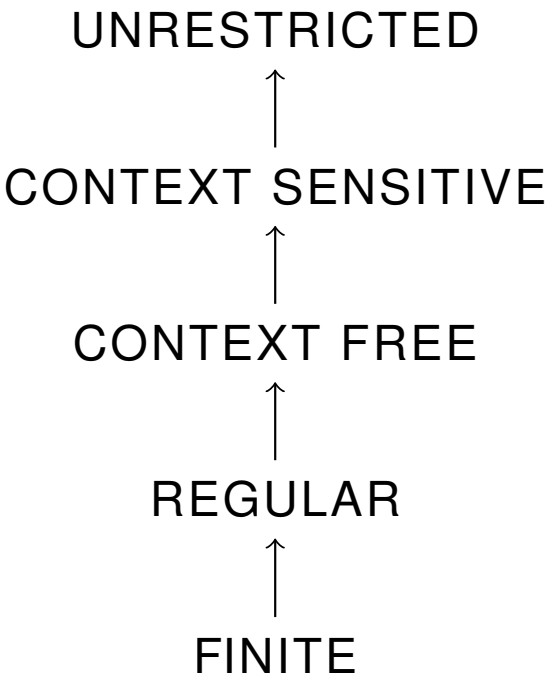
... das mer d'chind em Hans es huus lönd hälfe aastriiche
... that we the children-ACC Hans-DAT house-ACC let help paint



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

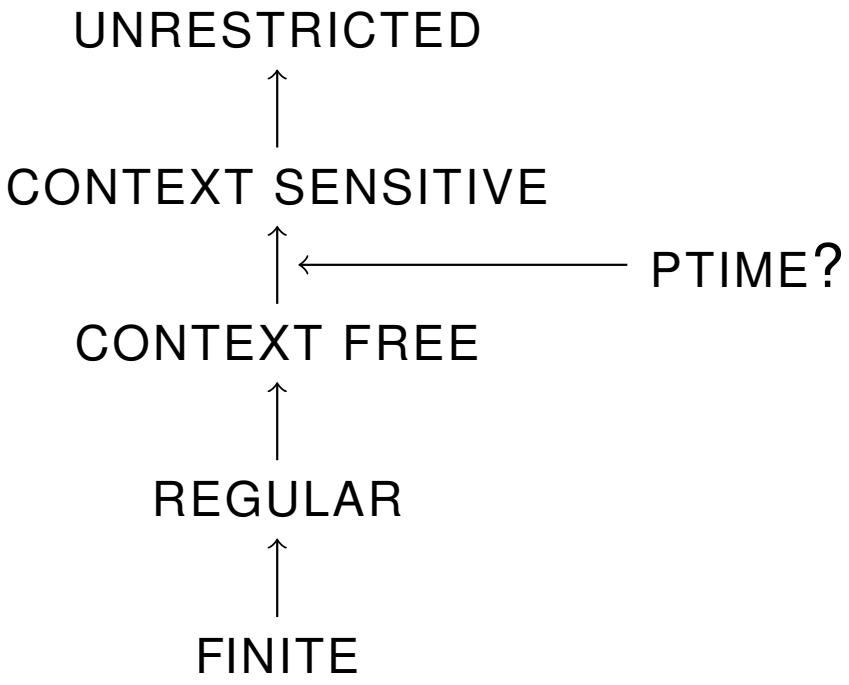
Chomsky hierarchy

Top down



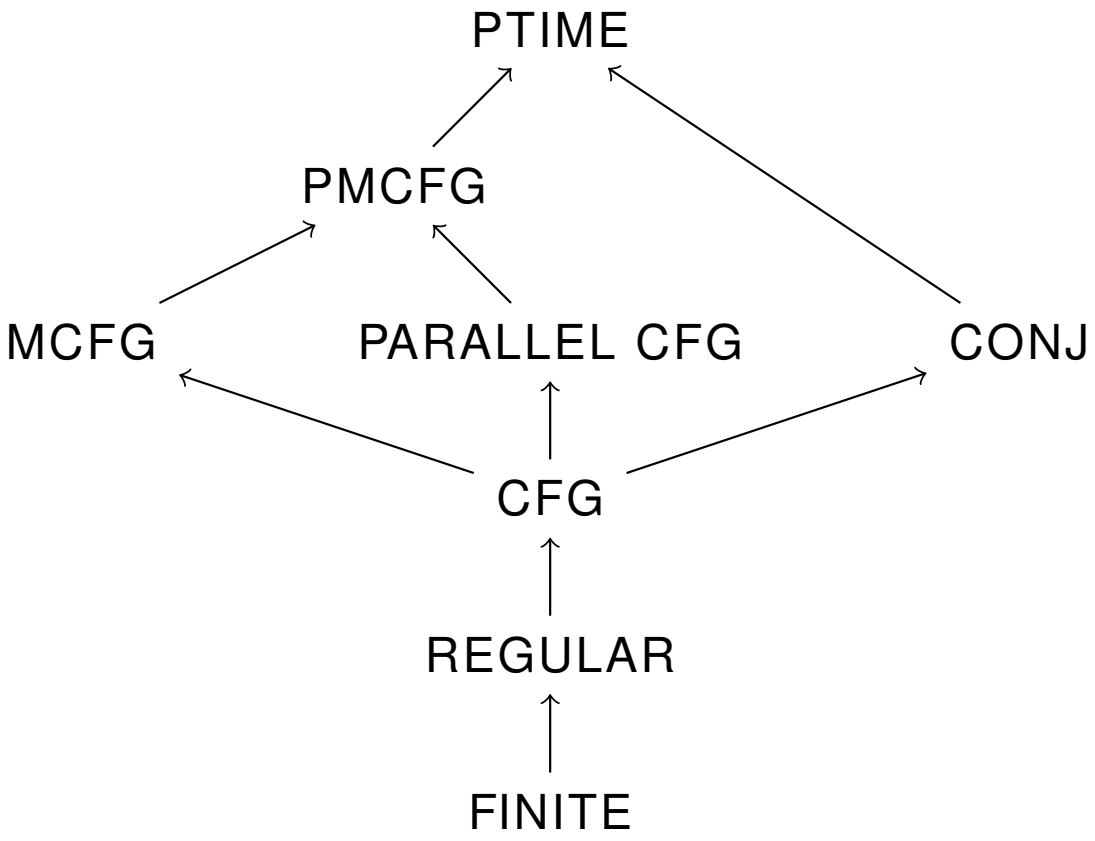
Chomsky hierarchy

Top down



Chomsky hierarchy

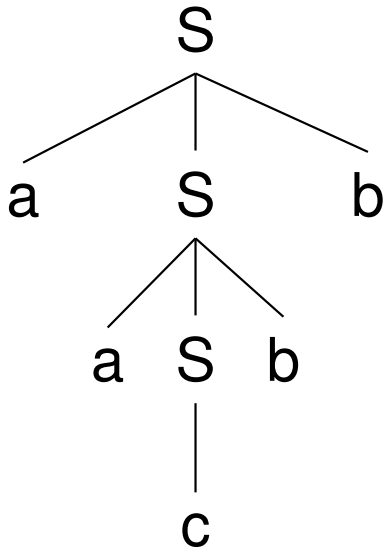
Bottom up



Outline

Derivation tree

Top-down versus Bottom-up



Top down versus bottom up

Kanazawa (2011)

Top down

Chomsky (1956, 1959)

- $N \rightarrow PQ$
- Rewrite N as PQ

Bottom up derivations

Smullyan (1961), Chomsky (1995), Stabler (1997)

- $N \leftarrow PQ$
- Combine P and Q to make N
- $N(xy) := P(x), Q(y)$

Minimalist program and MERGE

External merge

Combine X and Y to make a new object.

Internal merge

Y is part of X

Minimalist Grammars (Stabler, 1997)

Tractable formalisation of this approach.

Mildly context sensitive languages

Joshi

Slightly vague and informal definition:

Definition

A class of languages is MCS if:

- Not too complex: Efficiently parsable (in P)
- Constant growth property/semilinearity
- Includes some classes of crossing dependencies
- (includes all context-free languages)

Convergence of MCS formalisms

Joshi et al. (1990)

A crucial result for linguistics: all of the then current proposals turned out to be weakly equivalent:

- Linear indexed grammars
- Tree adjoining grammars
- Head grammars
- Combinatory categorial grammars

The derivation structures can all be described using LCFRs (we will use Multiple Context Free Grammars instead)

Three interesting classes

- Multiple context free grammars
- Well-nested Multiple Context-Free Grammars
- Well-nested Multiple Context-Free Grammars of dimension 2

Minimalist Grammars

Equivalence of MCFGs and Minimalist Grammars

MGs are weakly and strongly equivalent to MCFGs
Derivation trees of MGs can be learned.
Derived trees can be deterministically generated from the derivation trees
This gives a natural treatment of “movement”

Consensus

[Stabler(2013)]

This consensus is stable and rather well understood

- Broad consensus that somewhere in the MCFG hierarchy is adequate for syntax.
- Some problematic cases but all a bit marginal
- Possibly we need some true copying operation as well
  $N_1(uu) \leftarrow P_1(u)$
  Also for full stem reduplication in morphology?
- Alternatively there might be some nontrivial operations at Spellout.

Quick guide to MCFGs

Notation for CFG

$$N \to aPQ$$

Quick guide to MCFGs

Notation for CFG

$$N \to aPQ$$

$$N(auv) \leftarrow P(u)Q(v)$$

Quick guide to MCFGs

Notation for CFG

$$N \to aPQ$$

$$N(auv) \leftarrow P(u)Q(v)$$

$$N(auv) \leftarrow Q(v)P(u)$$

MCFG rules

Some nonterminals can have dimension 2:

$$N_2 \stackrel{*}{\Rightarrow} \langle \text{which pictures of himself, disliked} \rangle$$

$$N_2(u_1 v_1, u_2 a v_2) \leftarrow P_2(u_1, u_2) Q_2(v_1, v_2)$$

Outline

Three relations

Regular

$l \sim r$ iff $lr \in L$

Context-substring

$(l, r) \sim u$ iff $lur \in L$

Three relations

Regular

$l \sim r$ iff $lr \in L$

Context-substring

$(l, r) \sim u$ iff $lur \in L$

Natural generalisation

$(l, m, r) \sim (u, v)$ iff $lumvr \in L$

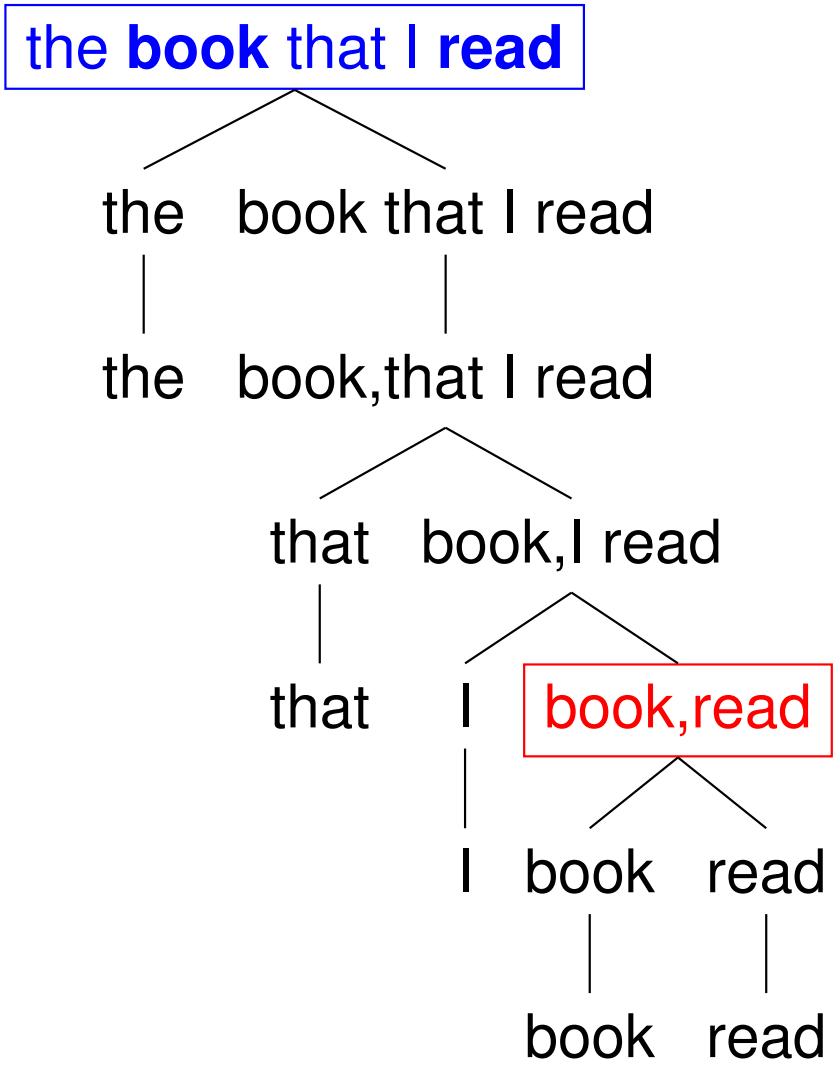
Multi-contexts

[Yoshinaka(2011)]

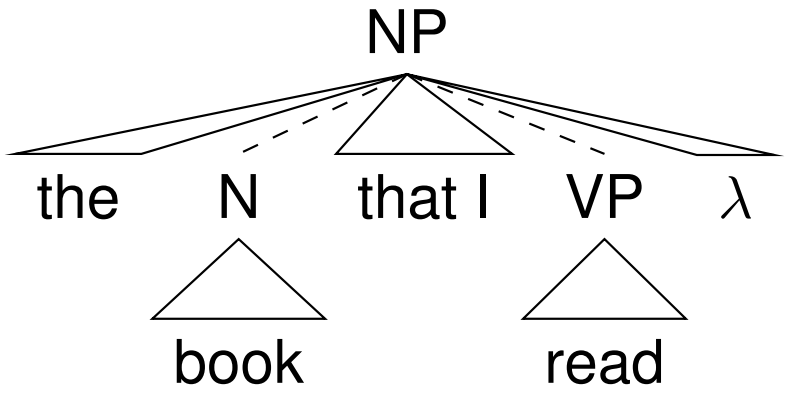
"this is the book that I told you to read"

- "this is □ that I told you to □ "
- $\langle$the book, read$\rangle$

Extended contexts



Extended contexts



Example

$L = \{cwcw | w \in (a, b)^*\}$

$\langle c, c \rangle \equiv_L \langle ca, ca \rangle \equiv_L \langle cbb, cbb \rangle$

Rules

Combine $\langle c, c \rangle$ with $\langle a, a \rangle$ to get $\langle ca, ca \rangle$

Converts $\langle cw, cw \rangle$ to $cwcw$

[Clark and Yoshinaka(2013)]

Theorem

If every nonterminal of dimension d in a MCFG can be represented by a finite set of contexts of dimension d, then we can learn the language.

[Clark and Yoshinaka(2013)]

Theorem

If every nonterminal of dimension d in a MCFG can be represented by a finite set of contexts of dimension d, then we can learn the language.

Caveats

- Membership queries
- Weak result
- Too slow as MCFGs are bigger than MGs.

Outline

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.

Linguistic examples

- Reduplication in morphology and phonology
- Suffixaufnahme in Old Georgian
- Yoruba copying
- Chinese number names

Reduplication

- Unbounded full stem reduplication in morphology (Inkelas, 2008)
- Washo, Malay, Dyirbal, . . .

Dyirbal nominals

Singular	Plural
midi	midi-midi
gulgiri	gulgiri-gulgiri

Yoruba relative clauses

Kobele, 2006

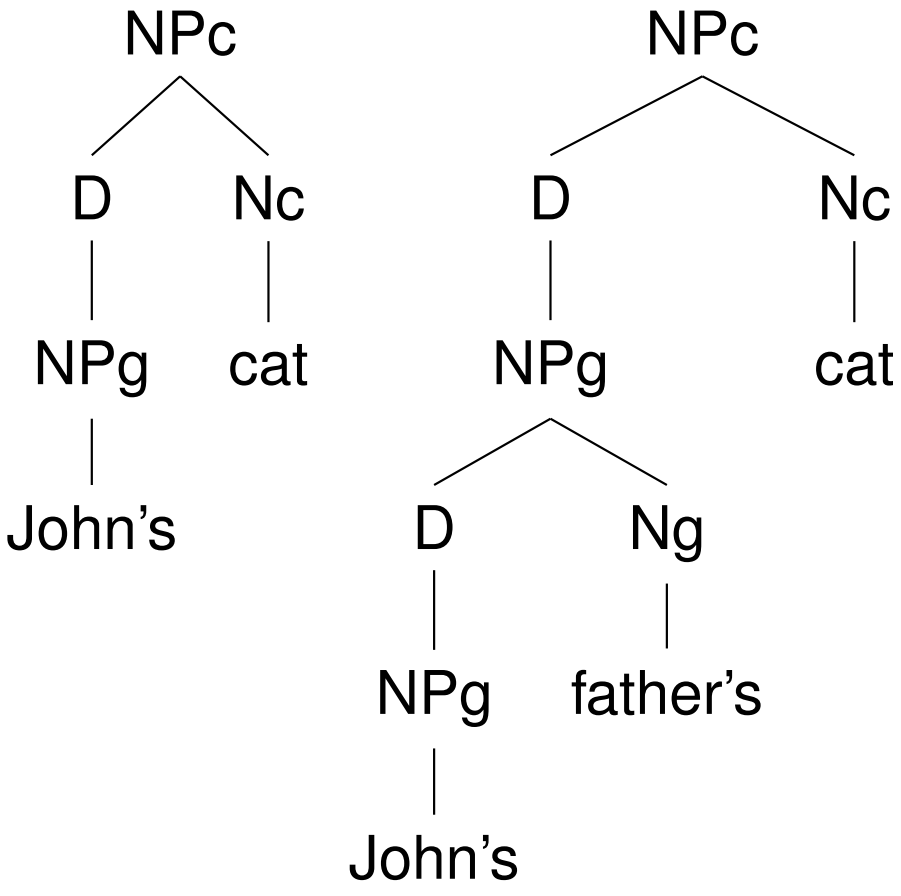
Recursive copying in Yoruba

'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.

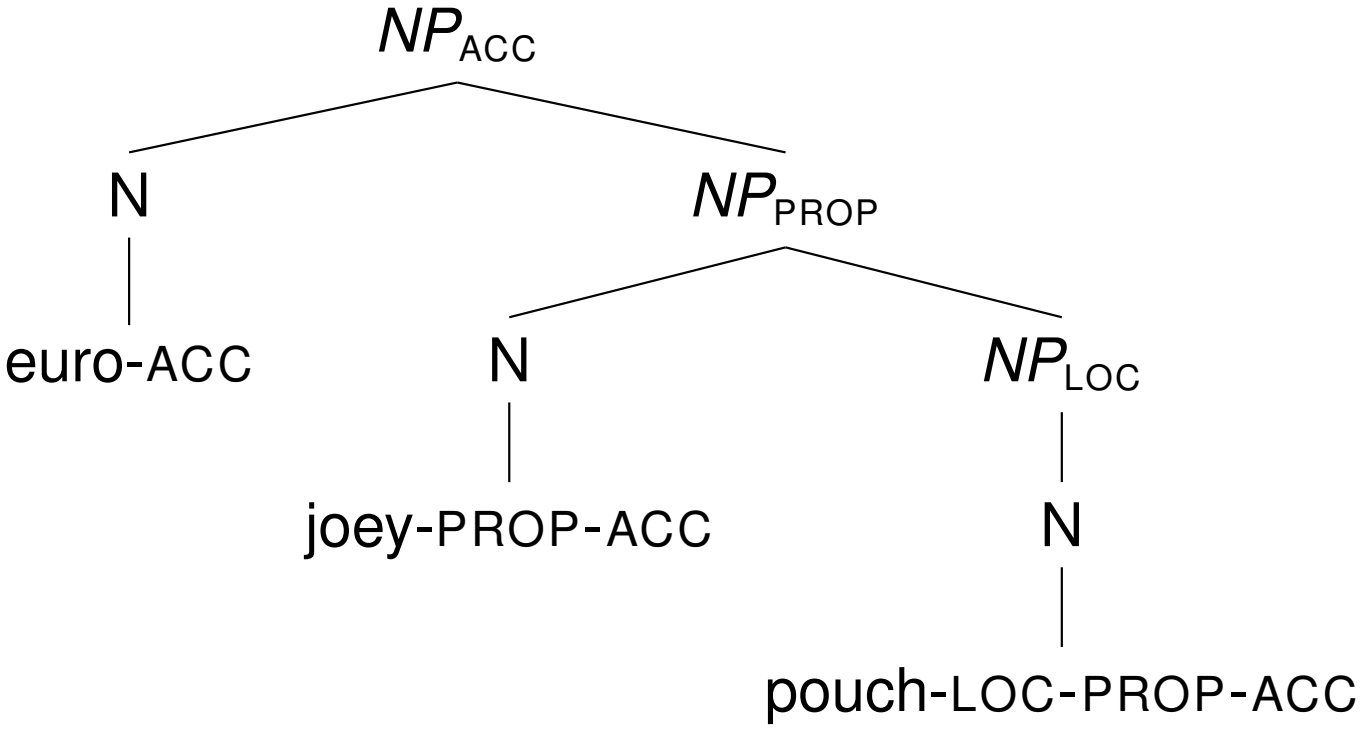
Suffixaufnahme/Case stacking

- 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

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

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

Parallel MCFGs

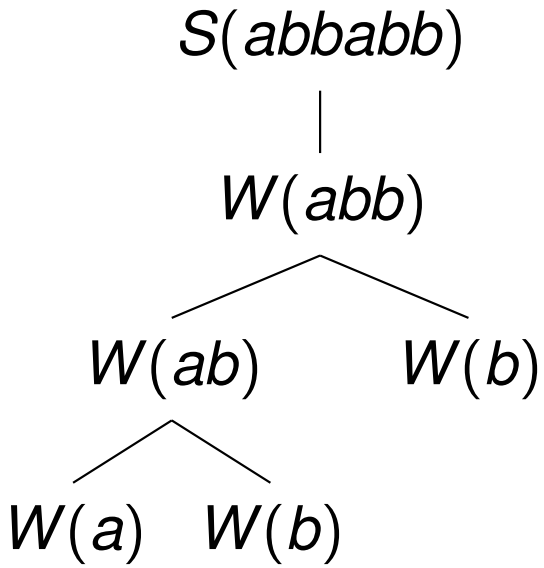
Copy language (semilinear)

$\{ ww \mid w \in \{a, b\}^+ \}$.

Simple PMCFG

$W(a) :-$
$W(b) :-$
$W(x_1 x_2) :- W(x_1), W(x_2)$
$S(x_1 x_1) :- W(x_1)$

Bottom up derivations



- $\vdash_G W(abb)$
- $S(x_1 x_1) :- W(x_1)$
- So applying this rule with $x_1 = abb$, we get $\vdash_G S(abbabb)$

Trivial example of a non-semilinear language

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

$S(x_1 x_1) :- S(x_1)$

$S(a) :-$

$S(aaaaaaaa)$

|

$S(aaaa)$

|

$S(aa)$

|

$S(a)$

$S(nngngggv)$

|

$N(nngn, gg)$

|

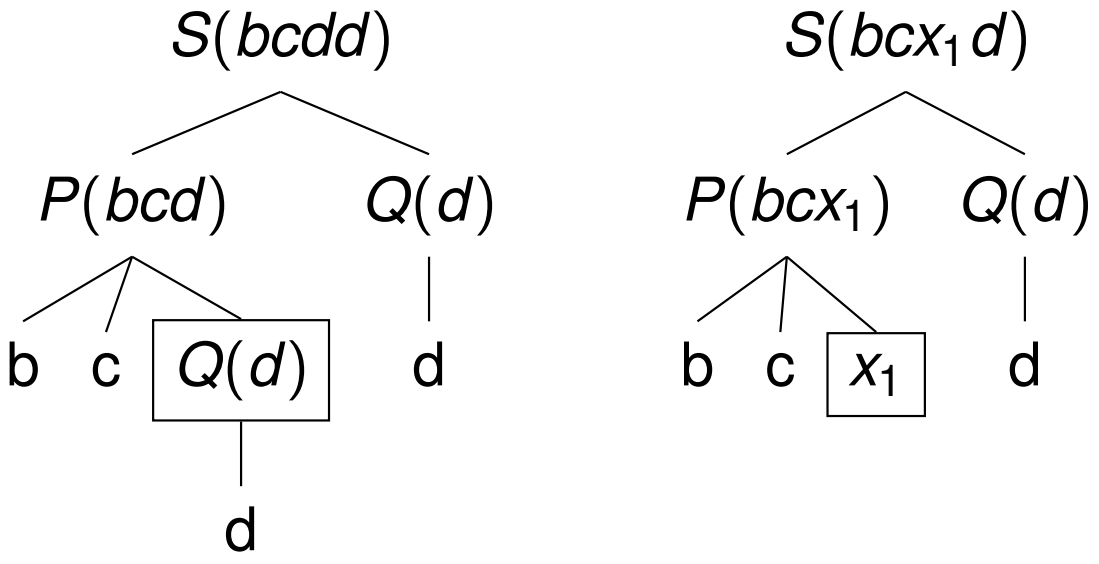
$N(nn, g)$

|

$N(n, \lambda)$

$$S(x_1 x_2 v) :- N(x_1, x_2) \ ;$$
$$N(x_1 x_2 n, x_2 g) :- N(x_1, x_2) \ ;$$
$$N(n, \lambda) :- \ .$$

Context in a CFG derivation

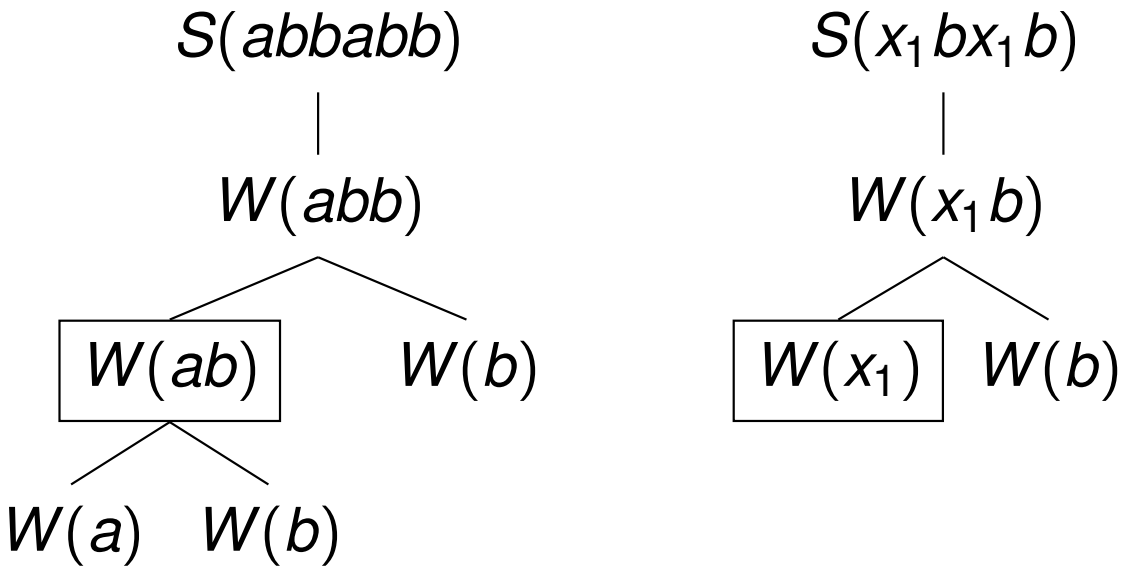


Function

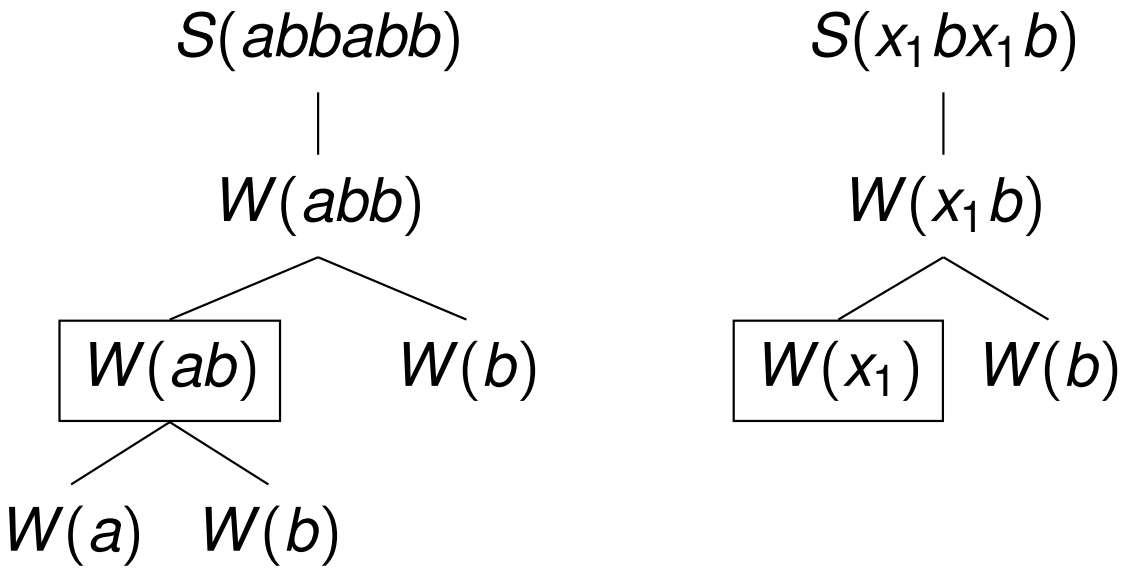
$f(x_1) = bcx_1d$

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

Generalised Context



Generalised Context



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

Target class

Target class

$\mathbb{G}(p, q, r, s)$
Fix p, q, r, s to be small (1 or 2 . . .)

- p dimension and q rank (standard MCFG hierarchy)
- r copying
- s number of contexts

Examples

All regular languages are in $\mathbb{L}(1, 1, 1, 1)$
Dyck language is in $\mathbb{L}(1, 2, 1, 1)$
Copy language is in $\mathbb{L}(1, 1, 2, 1)$

There are CFLs that are not in this class for any p, q, r, s

Conclusion

- A learning result for PMCFGs using a dual representation.
- Primal algorithms seem not to work at this level.
- Corollary: a dual result for MCFGs.
- An efficient weak learning result for a class that plausibly includes all natural languages.

Further results

These results can be generalised to learning other types of problem:

Yoshinaka and Kanazawa, LACL 2011

3 results for abstract categorial grammars

Yoshinaka and Kasprzik, DLT 2011

Learning context free tree languages

Clark ICML 2011

Learning context free transducers from input output pairs

Outline

Weak and strong learning

Weak learning

Class of languages is probably weakly adequate:
What subset of techniques are actually used?
Unlearnable languages: $\{a^n b^m | n \neq m\}$

Strong learning

Much more limited $\Rightarrow$ stronger predictions
Only one structural mapping per weak language

Some strong predictions

- For a word u let $Lex(u)$ be the set of lexical entries for u.
- $C_L(u)$ be the distribution of u

Strong identity implies weak identity

If $Lex(u) = Lex(v)$ then $C_L(u) = C_L(v)$

Some strong predictions

- For a word u let $Lex(u)$ be the set of lexical entries for u.
- $C_L(u)$ be the distribution of u

Strong identity implies weak identity

If $Lex(u) = Lex(v)$ then $C_L(u) = C_L(v)$

Prediction: weak identity must imply strong identity

If $C_L(u) = C_L(v)$ then $Lex(u) = Lex(v)$

Some strong predictions

- For a word u let $Lex(u)$ be the set of lexical entries for u.
- $C_L(u)$ be the distribution of u

Strong identity implies weak identity

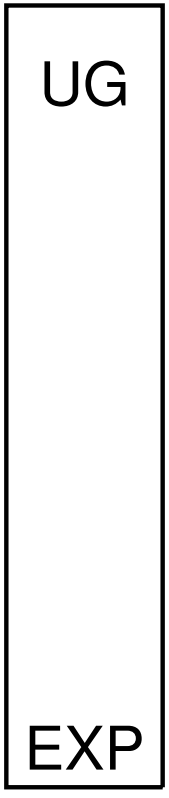
If $Lex(u) = Lex(v)$ then $C_L(u) = C_L(v)$

Prediction: weak identity must imply strong identity

If $C_L(u) = C_L(v)$ then $Lex(u) = Lex(v)$
If $C_L(u) \supseteq C_L(v)$ then $Lex(u) \supseteq Lex(v)$

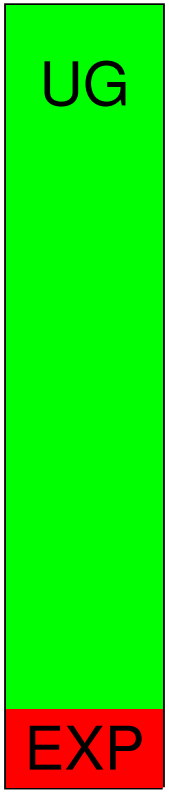
Where does language come from?

Two factors: UG and experience



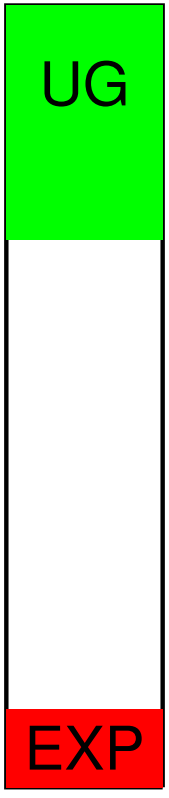
Where does language come from?

Principles and Parameters



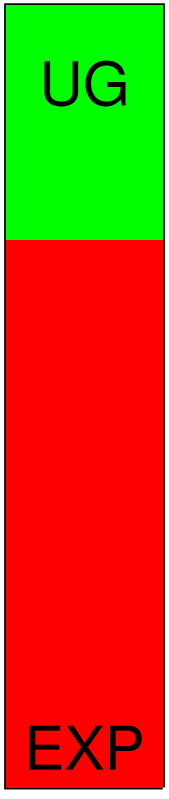
Where does language come from?

MP/Biolinguistics



Where does language come from?

Distributional learning



Three factors

Chomsky (2005)

"1. Genetic endowment, apparently nearly uniform for the species, which interprets part of the environment as linguistic experience, a nontrivial task that the infant carries out reflexively, and which determines the general course of the development of the language faculty. . . .
2. Experience, which leads to variation, within a fairly narrow range, as in the case of other subsystems of the human capacity and the organism generally.
3. Principles not specific to the faculty of language."

Third factor principles

Chomsky, 2005/2006

"The third factor falls into several subtypes: (a) principles of data analysis that might be used in language acquisition and other domains; (b) principles of structural architecture and developmental constraints that enter into canalization, organic form, and action over a wide range, including principles of efficient computation, which would be expected to be of particular significance for computational systems such as language. It is the second of these subcategories that should be of particular significance in determining the nature of attainable languages."

"It also might turn out that general cognitive principles that enter into language acquisition pose conditions on FL design."

Questions

We have a mathematical theory of how rich grammars might be learned from strings by a process of inductive generalisation.

- Can we fill in the missing pieces of the theory?
- Can we turn this into a theory of language acquisition?
- Can we learn what linguists are interested in?
- How much of generative grammar survives?

Bibliography I


Alexander Clark and Ryo Yoshinaka.
Distributional learning of parallel multiple context-free grammars.
Machine Learning, pages 1–27, 2013.
ISSN 0885-6125.
doi: 10.1007/s10994-013-5403-2.
URL
http://dx.doi.org/10.1007/s10994-013-5403-2.


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

Bibliography II


R. Yoshinaka.
Efficient learning of multiple context-free languages with multidimensional substitutability from positive data.
Theoretical Computer Science, 412(19):1821 – 1831, 2011.