Factored grammar and performance models

Edward Stabler, UCLA

ENS Paris 2012

- Factors in grammars and performance
  - ◦ MG (merge,move) vs MCFG ($\rightarrow$): strongly $\equiv$ but $\neq$
  - $*$ Factors in incremental parsing
  - ◦ Another factor: MG+ϕAgree vs MG
  - ◦ Nonissue: Traces vs none
- Certain varieties of structure dependence matter:
  how to defend these claims

Computational models beyond level 1: Basics first

(1) All relevant responses in range. "descriptive adequacy, level 1"

We need not have the correct model; but a class containing it.

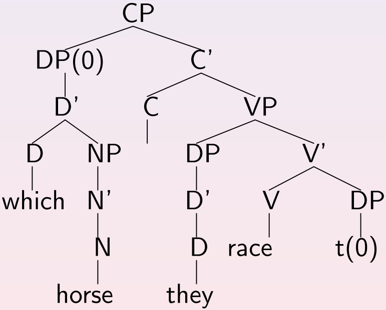
(2) Among adequate models, how to choose?

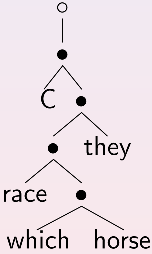
- $\mathcal{O}(n)$ differences in space/time are insignificant.

Equivalents with a few symbols more or less mainly uninteresting.

- $\mathcal{O}(2^n)$ or $\mathcal{O}(2^{\cdot^{\cdot^{2^n}}})$ differences significant.

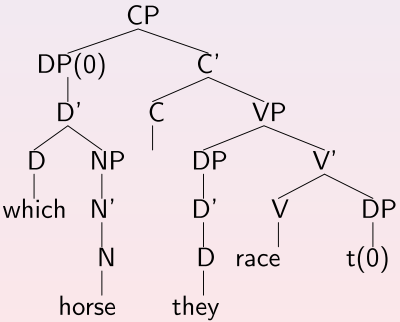
In/significant comparisons confused in literature; contrasted here.


CP
DP(0)
C'
D'
C
VP
D
NP
DP
V'
which
N'
D'
V
DP
N
D
race
t(0)
horse
they

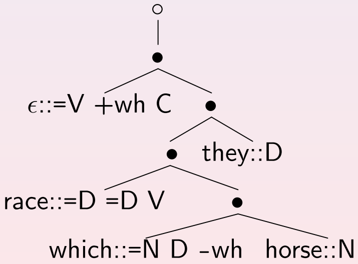

C
they
race
which
horse

Minimalist grammars (MGs)

derived tree



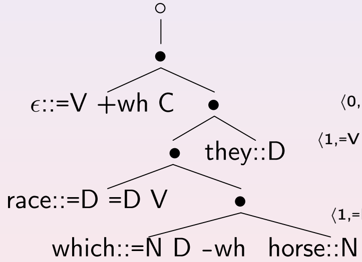
derivation tree

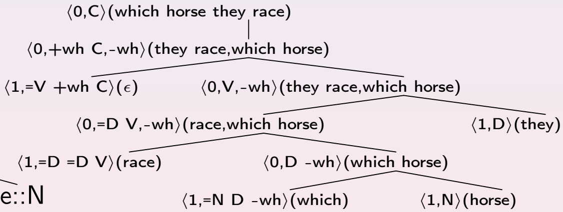


⇑

isomorphic to MCFG derivation

MGs ≡ MCFGs

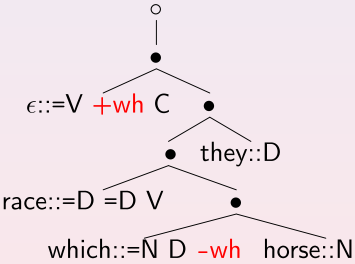


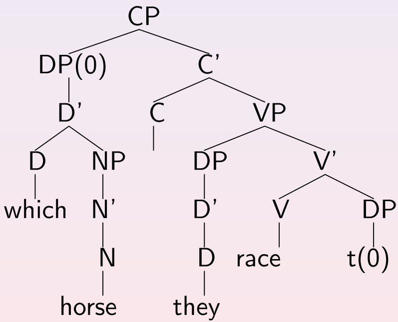


- MG features treated as MCFG categories: relation is transparent!
- This translation always works – every MG strongly equiv to MCFG

(Michaelis'98,'01; Harkema'01)

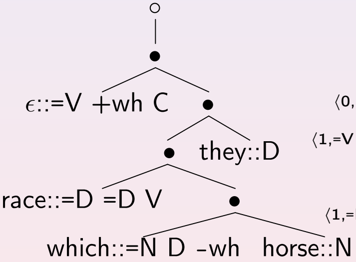
Movement in MG

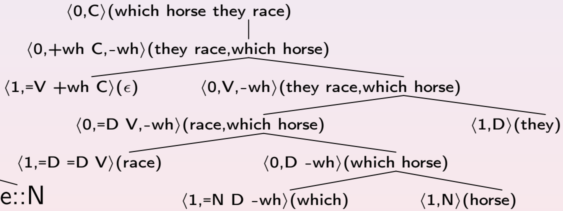




$$\frac{s : +f\gamma, \mu \uplus \{t : \text{-}f\}}{ts : \gamma, \mu}\ (\circ_1)$$

Movement in MCFG





$$\langle 0,\text{=D V,-wh}\rangle(x,y) \rightarrow \langle 1,\text{=D =D V}\rangle(x) \quad \langle 0,\text{D -wh}\rangle(y)$$
$$\langle 0,\text{C}\rangle(yx) \rightarrow \langle 0,\text{+wh C,-wh}\rangle(x,y)$$

MG vs MCFG movement: significantly different

- MG treats movement configurations [+$f\alpha$] . . . [-$f\beta$] . . . alike, but MCFG needs a separate rule for every instance

- This allows us to prove: MGs can be exponentially smaller than strongly equivalent MCFGs.

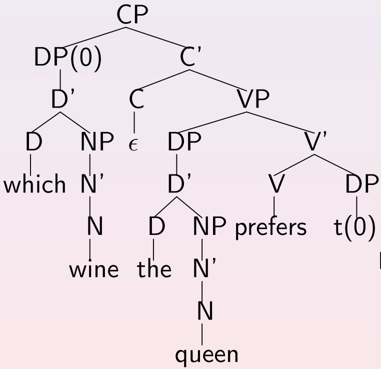
  For any k we show how to define MG with k movers that can be introduced to a XP in any order; any equivalent MCFG needs at least 2^k rules. $\square$

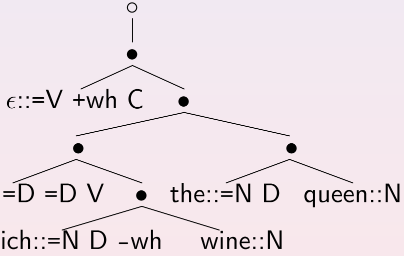
(pf)

(M)CFG explanatory inadequacy

- I do not know whether English is. . . literally beyond the bounds of phrase structure description. . . When we turn to the question of the complexity of description. . . , however, we find that there are ample grounds for the conclusion that this theory of linguistic structure is fundamentally inadequate. (Chomsky'56, p.119)
   - Pullum: "[A weak] non-CF-ness result itself, Chomsky has repeatedly told us, is of little importance."
- Does the idea that Gs should distinguish movements across categorial differences have a bearing on performance models? Is it supported by evidence from performance?

Performance models

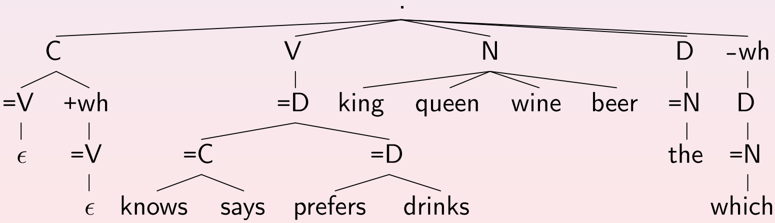




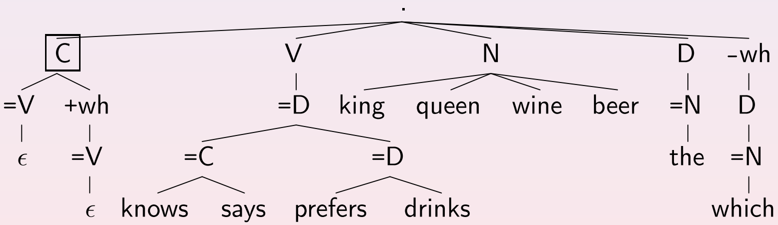
- In most models, grammar size $\propto$ parser size.
- Often both grammar + ops explicit: 1 step/node in derivation...

Model: TD beam parser for MGs

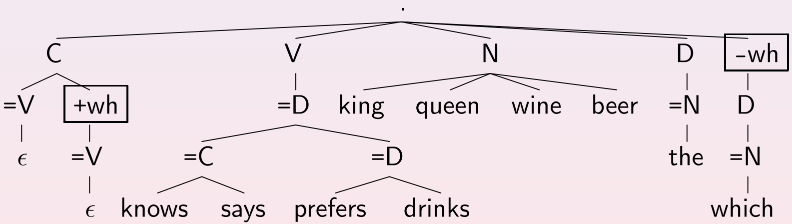
ϵ::=V C	knows::=C =D V	king::N	the::=N D
ϵ::=V +wh C	says::=C =D V	queen::N	which::=N D -wh
	prefers::=D =D V	wine::N	
	drinks::=D =D V	beer::N	



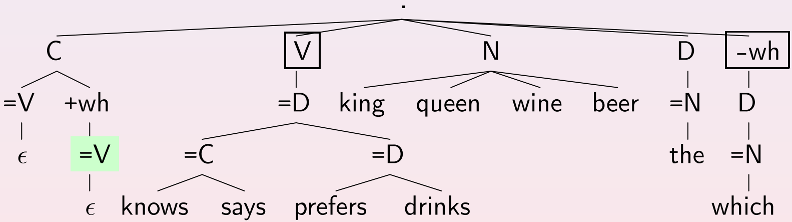
step	remaining input	rule	queue
0.	which wine the queen prefers	start	$\boxed{1}$



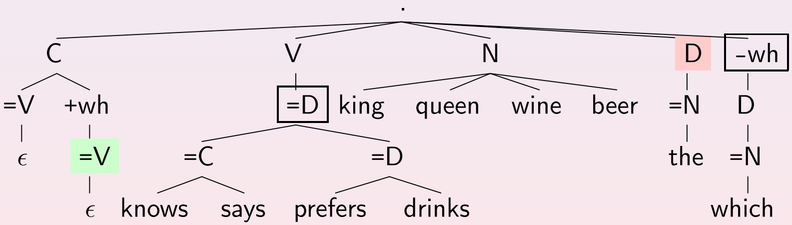
step	remaining input	rule queue
1.	which wine the queen prefers	$\circ_1$ $\boxed{1}$



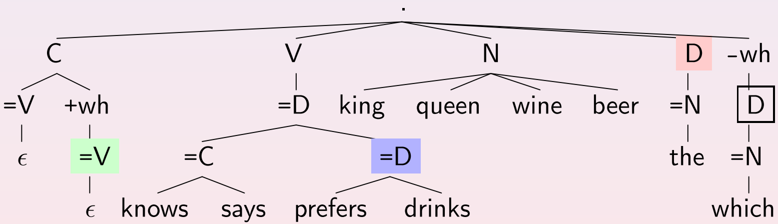
step	remaining input	rule	queue
2.	which wine the queen prefers	$\bullet_1$	$\boxed{1}$ 2



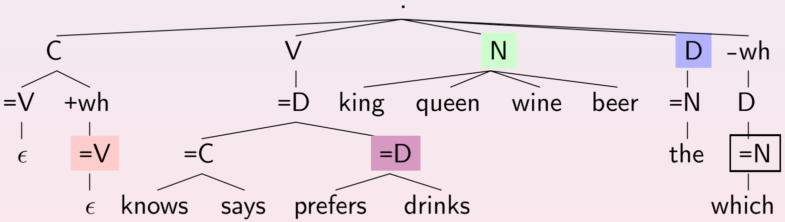
step	remaining input	rule queue
3.	which wine the queen prefers	$\bullet_2$ $\boxed{1}$ 2 3



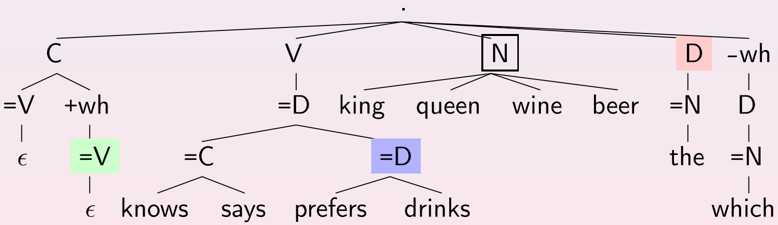
step	remaining input	rule queue
4.	which wine the queen prefers	$\bullet_3$ $\boxed{1}$ 2 3 4



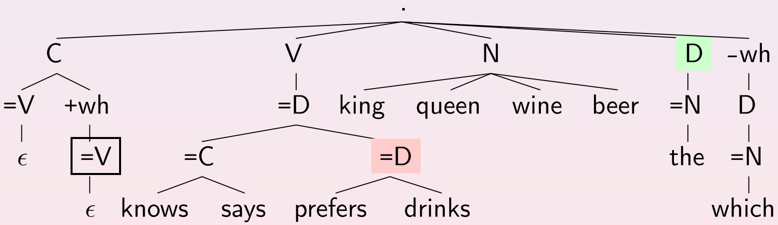
step	remaining input	rule	queue
5.	which wine the queen prefers	$\bullet_1$	$\boxed{1}$ 2 3 4 5



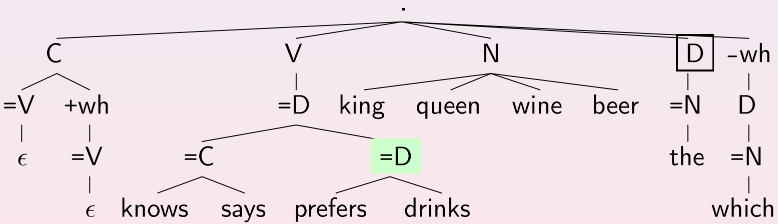
step	remaining input	rule	queue
6.	~~which~~ wine the queen prefers	scan	$\boxed{1}$ 2 3 4



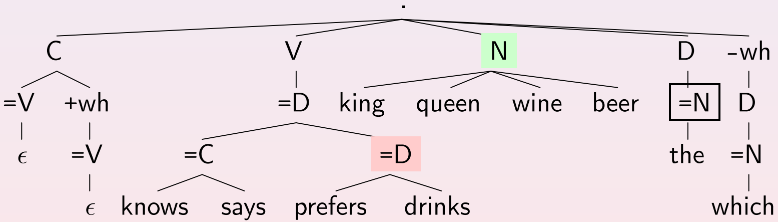
step	remaining input	rule	queue
7.	~~which wine~~ the queen prefers	scan	$\boxed{1}$ 2 3



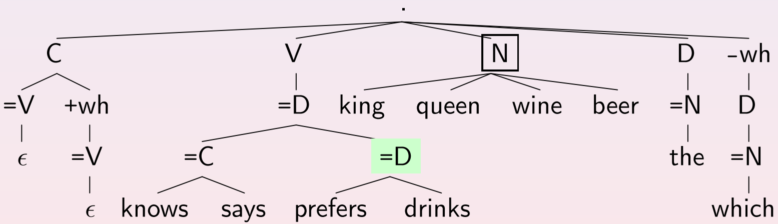
step	remaining input	rule	queue
8.	~~which wine~~ the queen prefers	scan	$\boxed{1}$ 2



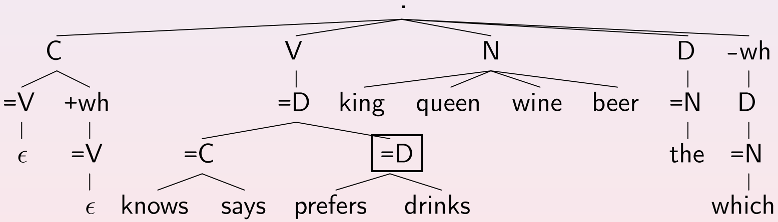
step	remaining input	rule	queue
9.	~~which wine~~ the queen prefers	$\bullet_1$	$\boxed{1}$ 2 3



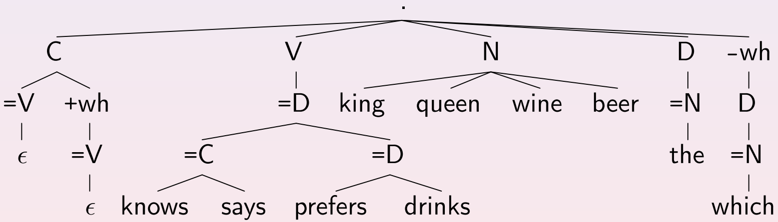
step	remaining input	rule	queue
10.	~~which wine the~~ queen prefers	scan	$\boxed{1}$ 2



step	remaining input	rule	queue
11.	~~which wine the queen~~ prefers	scan	$\boxed{1}$



step	remaining input	rule	queue
12.	~~which wine the queen prefers~~	scan	ϵ

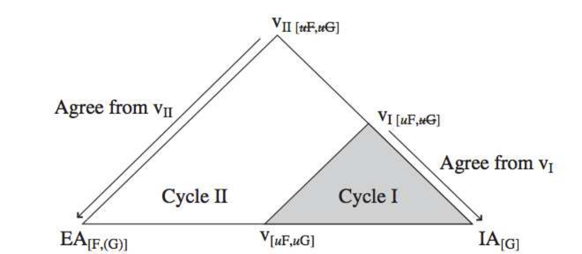


- this method works for any MG (sound, complete)
- MG representation is transparent; succinctness evident

Independent evidence: movement distinguished

- **locality effects in movement:** ∘ self-paced reading effect $\propto$ distance to antecedent (Gibson'98;Hale'03;Bartek&al'11,...)
  ∘ island effects (Aoshima&al'09,Sag&al'07,Yoshida'06,...)
- **patterns of acquisition:** aquisition of wh-movement, etc. (Friedmann&Lavi'06, vanKampen97')
- **neural correlates?** Brodman 45 for movement (Santi&Grodzinsky'10)

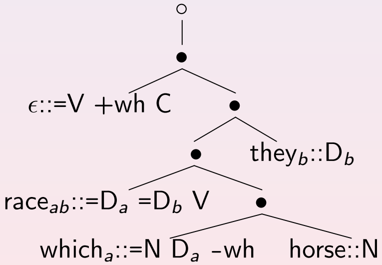
ϕ agreement



Agree domain: Bejar & Rezac 2009

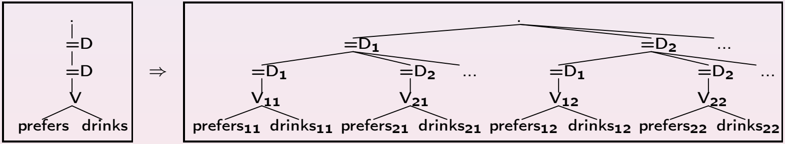
Kobele'11: Many proposals of this kind are regular constraints:
Graf'11: enforceable in the MG category system

ϕ agreement in MGs



How to allow a derivation like this for each ϕ specification a, b that the language allows?

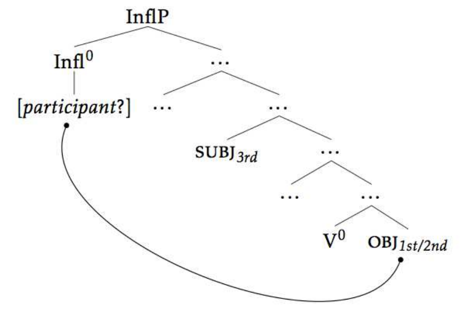
ϕ agreement in MGs



If DP has k features with j values, then j^k possibilities.
If n args agree, 'categorial infrastructure' multiplied by 2^{kn},
missing generalizations about match of verb+arguments.

(pf)

ϕAgr probes



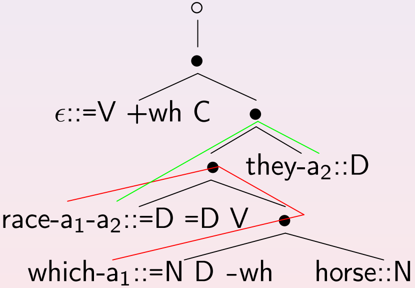
Preminger 2011

- suppose each arg has its own probe
- then instead of j^{nk} grammar, we have nj^k

(U1) perhaps $n \leq 2\frac{1}{2}$?

(U2) perhaps j, k bounded too? still j^{nk} can be large

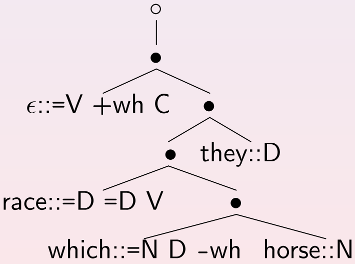
MG+ϕAgr



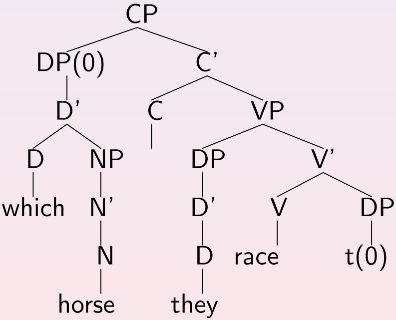
Each ϕAgr probe realized by regular tree automaton

States of each ϕAgree probe not multiplied through Lex, parser smaller, generalization captured.

(D) derivation tree

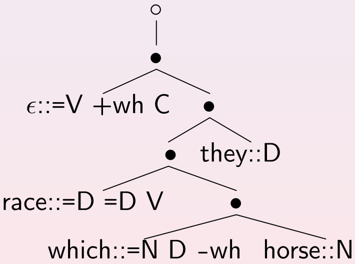


(T) derived tree (traces)

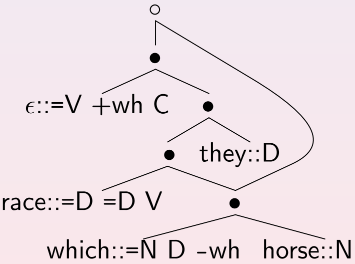


Insignificantly different alternatives

(D) derivation tree



(M) derived tree (multidominance)



D, T, M are easily, $\mathcal{O}(n)$ computable from each other.
(Graf'11;Kobele'11;Kobele&al'07;Mönnich'07)

Varieties of structure dependence

> [subj-aux inv] refers to the abstract label "noun phrase," a grouping of words into constituents, and consequently is called *structure dependent -Berwick,Pietroski,Yankama,Chomsky'11*

- Rules of Gs here define constituents, 'groupings of words' like NP
- MG rules 'groups constituents' that trigger movement, and MG+ϕAgree 'groups constituents' relevant to agreement.

$\Rightarrow$ non-construction-specific, more succinct, captures generalizations.

> Stipulation of grammatical constructions (interrogative, passive, etc.), with their independent properties, was overcome . . . by analyzing them into components that function generally, also eliminating redundancies – *Chomsky'12*

Questions about models and methods

Q1 How should models be developed and empirically assessed?

* Desc adequacy sets stage for next questions

  Math. characterization of classes of possibilities!

Q2 Most significant gaps in our understanding?

q1 What are the basic mechanisms of grammar?

  Beyond level 1: what is basic, how to defend?

  What performance measures bear most directly on grammar?

q2 How does the language learner work?

Q3 Experimental approaches vs computational modeling??

* q1 assessment difficult. cf. Newton/Hooke, field/math bio

APPENDICES

A. Relative succinctness of MGs vs. MCFGs

For each $i \in \mathbb{N}$, the lexicon of $\mathrm{MG_i}$ contains the lexical items specified in (i-iv) and nothing else, with A the 'start' category:

i. the following lexical item, with $i + 2$ syntactic features:

$$\mathrm{a} :: \mathrm{=B\ +1+2\ldots+i\ A}$$

ii. the following lexical item, with 3 syntactic features:

$$\mathrm{b} :: \mathrm{=B=C\ B}$$

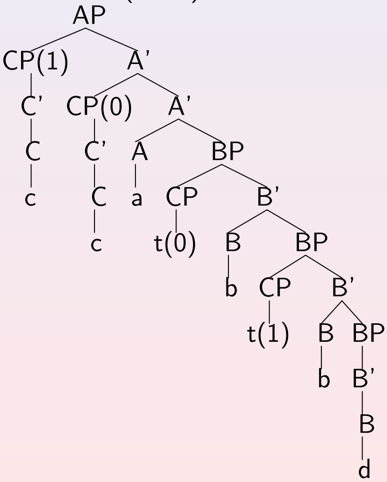
iii. the following lexical item, with 1 syntactic feature:

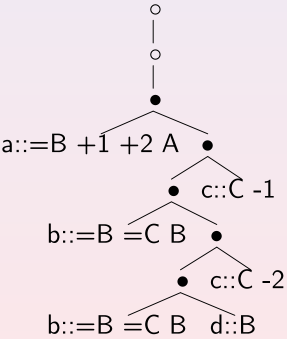
$$\mathrm{d} :: \mathrm{B}$$

iv. And for each $1 \leq j \leq i$, this lexical item with 2 syntactic features,

$$\mathrm{c} :: \mathrm{=C\ -j}.$$

MG_2 has $(3*2)+6=12$ features; 2! derivations of $\mathrm{c}^2\mathrm{ab}^2\mathrm{d}$:





B. Relative succinctness of MG+ϕAgr vs. MG

We define a series of MGs, MG$_i$ for $i = 1, 2, \ldots$ For simplicity, assume each argument is a pronoun $k = 1$ feature with $j = 2$ values, 0 or 1; each verb in MG_n selects and agrees with n arguments. So MG$_n$ needs, for $j, j_i \in \{0, 1\}$,

$$\text{pronoun-}j\text{::D}_j$$
$$\text{verb}-j_n \ldots - j_1\text{::=D}_{j_n} \ldots \text{=D}_{j_1}\ \text{V}$$

That is $2 + 2^n$ lexical items in each MG$_n$.
For comparison, we can use probes to define the same languages. 2 lexical items define what Baker (2001) calls ‘categorial infrastructure’:

pronoun::D
verb::=D. . . =D V

For each of the n arguments of V, an automaton (probe) matches an affix 0 or 1 with the corresponding affix of V. (Assume verb,pronoun dominate their affixes, so the automation can find them.) That’s 2 lexical items and a collection of n automata each of which has a size $\mathcal{O}(n)$.

C. Structure dependence vs. linear order

Chomsky (1968, 1971, 1975, 1980a,b) discusses 'structure dependence' in many places, but leaves it not quite clear. G is structure dependent iff it has rules that generate expressions from other expressions, so that it can, for example, have a rule that applies to 'groupings of words' like the 'noun phrases' to form other expressions like 'determiner phrases'. In this sense...

- 'Orthogonal to' nested, hierarchical structure?
  No. Grammars for nested structure are all structure dependent.
  vE.g. [[a] [b] [c]] => [[b] [[a] [c]]] refers to nested structures
- Distinct from 'linear structure'? What is that?
  'the *n*th verb' is definable even in a regular grammar.
- Hierarchical structure unable to count?
  No. Regular grammars can count to k; others arbitrarily high.

D. The TD beam parser for MG

Like CF beam parser except that stack is replaced by priority queue, sorted by linear precedence as explained in Stabler'11,'12.

Basically just inverting the standard bottom-up MG definitions, the parsing rules for MGs in Stabler'12, used in the example above:

$$\frac{}{\mathit{input},\ (\mathrm{C}(x),\emptyset)_\epsilon}\ \text{(start)}\ \ell[\mathrm{C}(x)],\ \text{for start category C}$$

$$\frac{w * \mathit{input},\ (t[w],\emptyset)_i * q}{\mathit{input},\ q}\ \text{(scan)}$$

$$\frac{\mathit{input},\ (t[{=}f(x)],\mu)_i * q}{\mathit{input},\ ({=}f(\Sigma x),\emptyset)_{i0} * (f(y),\mu)_{i1} * q}\ (\bullet_1)\ \ell[f(y)] \wedge \Sigma x \neq \epsilon$$

$$\frac{input,\ (t[{=}f(x)],\mu)_i * q}{input,\ ({=}f(\overline{\Sigma}x),\mu)_{i1} * (f(y),\emptyset)_{i0} * q}\ (\bullet_2)\ \ell[f(y)] \wedge \overline{\Sigma}x \neq \epsilon$$

$$\frac{input,\ (t[{=}f(x)],\ u[f(y)]_j \uplus \mu)_i * q}{input,\ ({=}f(\Sigma x),\emptyset)_i * (f(y),\mu)_j * q}\ (\bullet_3)\ \Sigma x \neq \epsilon$$

$$\frac{input,\ (t[{=}f(x)],\ u[f(y)]_j \uplus \mu)_i * q}{input,\ ({=}f(\overline{\Sigma}x),\mu)_i * (f(y),\emptyset)_j * q}\ (\bullet_4)\ \overline{\Sigma}x \neq \epsilon$$

$$\frac{input,\ (t[{+}f(x)],\mu)_i * q}{input,\ ({+}f(x),{-}f(y)_{i0} \uplus \mu)_{i1} * q}\ (\circ_1)\ \ell[{-}f(y)]$$

$$\frac{input,\ (t[{+}f(x)],\ u[{-}f(y)]_j * \mu)_i * q}{input,\ ({+}f(x),{-}f(y)_j \uplus \mu)_i * q}\ (\circ_2)$$

Bartek, B., Lewis, R. L., Vasishth, S., and Smith, M. R.
In search of on-line locality effects in sentence comprehension.
Journal of Experimental Psychology: Learning, Memory, and Cognition 37, 5 (2011), 1178–1198.

Barton, G. E., Berwick, R. C., and Ristad, E. S.
Computational Complexity and Natural Language.
MIT Press, Cambridge, Massachusetts, 1987.

Béjar, S., and Rezac, M.
Cyclic agree.
Linguistic Inquiry 40, 1 (2009), 35–73.

Chomsky, N.
Three models for the description of language.
IRE Transactions on Information Theory IT-2 (1956), 113–124.

Chomsky, N.
Language and Mind.
Harcourt Brace Javonovich, NY, 1968.

Chomsky, N.
Problems of Knowledge and Freedom: The Russell Lectures.
Vintage, NY, 1971.

Chomsky, N.
On cognitive structures and their development.
In *Language Learning: The Debate between Jean Piaget and Noam Chomsky*, M. Piattelli-Palmarini, Ed. Harvard University Press, Cambridge, Massachusetts, 1980, pp. 35–54.

Chomsky, N.
Problems of projection.

Lingua forthcoming (2012).

Culy, C.
The complexity of the vocabulary of Bambara.
Linguistics and Philosophy 8, 3 (1985), 345–352.

Friedmann, N., and Lavi, H.
On the order of acquisition of A-movement, wh-movement and V-C movement.
In *Language Acquisition and Development*, A. Belletti, E. Bennati, C. Chesi, E. D. Domenico, and I. Ferrari, Eds. Cambridge Scholars Press, Cambridge, 2006.

Gibson, E.
Linguistic complexity: Locality of syntactic dependencies.
Cognition 68 (1998), 1–76.

Graf, T.
Closure properties of minimalist derivation tree languages.
In *Logical Aspects of Computational Linguistics, LACL'11* (2011).

Graf, T.
Locality and the complexity of minimalist derivation tree languages.
In *Proceedings of the 16th Conference on Formal Grammar* (2011).
to appear.

Hale, J.
Grammar, Uncertainty, and Sentence Processing.
PhD thesis, Johns Hopkins University, 2003.

Harkema, H.
A characterization of minimalist languages.

In *Logical Aspects of Computational Linguistics* (NY, 2001), P. de Groote, G. Morrill, and C. Retoré, Eds., Lecture Notes in Artificial Intelligence, No. 2099, Springer, pp. 193–211.

Harkema, H.
Parsing Minimalist Languages.
PhD thesis, University of California, Los Angeles, 2001.

Kobele, G. M.
Minimalist tree languages are closed under intersection with recognizable tree languages.
In *Logical Aspects of Computational Linguistics, LACL'11* (2011).

Michaelis, J.
Derivational minimalism is mildly context-sensitive.
In *Proceedings, Logical Aspects of Computational Linguistics, LACL'98* (NY, 1998), Springer, pp. 179–198.

Michaelis, J.
On Formal Properties of Minimalist Grammars.
PhD thesis, Universität Potsdam, 2001.
Linguistics in Potsdam 13, Universitätsbibliothek, Potsdam, Germany.

Michaelis, J.
Transforming linear context free rewriting systems into minimalist grammars.
In *Logical Aspects of Computational Linguistics* (NY, 2001), P. de Groote, G. Morrill, and C. Retoré, Eds., Lecture Notes in Artificial Intelligence, No. 2099, Springer, pp. 228–244.

Preminger, O.
Agreement as a Fallible Operation.
PhD thesis, Massachusetts Institute of Technology, 2011.

Pullum, G. K.

Footloose and context-free.
Linguistic Inquiry 4, 3 (1986), 409–414.

Sag, I., Hofmeister, P., and Snider, N.
Processing complexity in subjacency violations: the complex noun phrase constraint.
In *Proceedings of 43rd Regional Meeting of the Chicago Linguistics Society* (2007).

Santi, A., and Grodzinsky, Y.
fMRI adaptation dissociates syntactic complexity dimensions.
NeuroImage 51, 51 (2010), 1285–1293.

Shieber, S. M.
Evidence against the context-freeness of natural language.
Linguistics and Philosophy 8, 3 (1985), 333–344.

Stabler, E. P.
Computational perspectives on minimalism.
In *Oxford Handbook of Minimalism*, C. Boeckx, Ed. Oxford University Press, Oxford, 2010, pp. 617–641.

Stabler, E. P.
After GB theory.
In *Handbook of Logic and Language, Second Edition*, J. van Benthem and A. ter Meulen, Eds. Elsevier, Amsterdam, 2011, pp. 395–414.

Stabler, E. P.
After GB theory.
In *Handbook of Logic and Language, Second Edition*, J. van Benthem and A. ter Meulen, Eds. Elsevier, Amsterdam, 2011, pp. 395–414.

Stabler, E. P.

Bayesian, minimalist, incremental syntactic analysis.
Topics in Cognitive Science forthcoming (2012).

van Kampen, J.
First steps in wh-movement.
PhD thesis, University of Utrecht, 1997.

Vijay-Shanker, K., Weir, D., and Joshi, A. K.
Characterizing structural descriptions produced by various grammatical formalisms.
In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics* (1987), pp. 104–111.

Yoshida, M.
Constraints and Mechanisms in Long-Distance Dependency Formation.
PhD thesis, University of Maryland, 2006.