Sharpening the empirical claims of generative syntax through formalization

Tim Hunter

University of Minnesota, Twin Cities

NASSLLI, June 2014

Part 1: Grammars and cognitive hypotheses
- What is a grammar?
- What can grammars do?
- Concrete illustration of a target: Surprisal

Parts 2–4: Assembling the pieces
- Minimalist Grammars (MGs)
- MGs and MCFGs
- Probabilities on MGs

Part 5: Learning and wrap-up
- Something slightly different: Learning model
- Recap and open questions

Sharpening the empirical claims of generative syntax
through formalization

Tim Hunter — NASSLLI, June 2014

Part 4

Probabilities on MG Derivations

Outline

13 Easy probabilities with context-free structure

14 Different frameworks

15 Problem #1 with the naive parametrization

16 Problem #2 with the naive parametrization

17 Solution: Faithfulness to MG operations

Outline

13 Easy probabilities with context-free structure

14 Different frameworks

15 Problem #1 with the naive parametrization

16 Problem #2 with the naive parametrization

17 Solution: Faithfulness to MG operations

Probabilistic CFGs

"What are the probabilities of the derivations?"

=

"What are the values of λ_1, λ_2, etc.?"

λ_1	S $\rightarrow$ NP VP
λ_2	NP $\rightarrow$ John
λ_3	NP $\rightarrow$ Mary
λ_4	VP $\rightarrow$ ran
λ_5	VP $\rightarrow$ V NP
λ_6	VP $\rightarrow$ V S
λ_7	V $\rightarrow$ believed
λ_8	V $\rightarrow$ knew

Probabilistic CFGs

"What are the probabilities of the derivations?"

=

"What are the values of λ_1, λ_2, etc.?"

λ_1	S → NP VP
λ_2	NP → John
λ_3	NP → Mary
λ_4	VP → ran
λ_5	VP → V NP
λ_6	VP → V S
λ_7	V → believed
λ_8	V → knew

→ Training algorithm (← Training corpus) →

1.0	S → NP VP
0.3	NP → John
0.7	NP → Mary
0.2	VP → ran
0.5	VP → V NP
0.3	VP → V S
0.4	V → believed
0.6	V → knew

$$\lambda_5 = \frac{\text{count}(\text{VP} \to \text{V NP})}{\text{count}(\text{VP})}$$

MCFG for an entire Minimalist Grammar

Lexical items:

$$\epsilon :: \langle \texttt{=t +wh c} \rangle_1 \qquad \text{praise} :: \langle \texttt{=d v} \rangle_1$$
$$\epsilon :: \langle \texttt{=t c} \rangle_1 \qquad \text{marie} :: \langle \texttt{d} \rangle_1$$
$$\text{will} :: \langle \texttt{=v =d t} \rangle_1 \qquad \text{pierre} :: \langle \texttt{d} \rangle_1$$
$$\text{often} :: \langle \texttt{=v v} \rangle_1 \qquad \text{who} :: \langle \texttt{d -wh} \rangle_1$$

Production rules:

$$\begin{aligned}
\langle st, u \rangle :: \langle \texttt{+wh c}, \texttt{-wh} \rangle_0 &\rightarrow s :: \langle \texttt{=t +wh c} \rangle_1 \quad \langle t, u \rangle :: \langle \texttt{t}, \texttt{-wh} \rangle_0 \\
st :: \langle \texttt{=d t} \rangle_0 &\rightarrow s :: \langle \texttt{=v =d t} \rangle_1 \quad t :: \langle \texttt{v} \rangle_0 \\
\langle st, u \rangle :: \langle \texttt{=d t}, \texttt{-wh} \rangle_0 &\rightarrow s :: \langle \texttt{=v =d t} \rangle_1 \quad \langle t, u \rangle :: \langle \texttt{v}, \texttt{-wh} \rangle_0 \\
ts :: \langle \texttt{c} \rangle_0 &\rightarrow \langle s, t \rangle :: \langle \texttt{+wh c}, \texttt{-wh} \rangle_0 \\
st :: \langle \texttt{c} \rangle_0 &\rightarrow s :: \langle \texttt{=t c} \rangle_1 \quad t :: \langle \texttt{t} \rangle_0 \\
ts :: \langle \texttt{t} \rangle_0 &\rightarrow s :: \langle \texttt{=d t} \rangle_0 \quad t :: \langle \texttt{d} \rangle_1 \\
\langle ts, u \rangle :: \langle \texttt{t}, \texttt{-wh} \rangle_0 &\rightarrow \langle s, u \rangle :: \langle \texttt{=d t}, \texttt{-wh} \rangle_0 \quad t :: \langle \texttt{d} \rangle_1 \\
st :: \langle \texttt{v} \rangle_0 &\rightarrow s :: \langle \texttt{=d v} \rangle_1 \quad t :: \langle \texttt{d} \rangle_1 \\
st :: \langle \texttt{v} \rangle_0 &\rightarrow s :: \langle \texttt{=v v} \rangle_1 \quad t :: \langle \texttt{v} \rangle_0 \\
\langle s, t \rangle :: \langle \texttt{v}, \texttt{-wh} \rangle_0 &\rightarrow s :: \langle \texttt{=d v} \rangle_1 \quad t :: \langle \texttt{d -wh} \rangle_1 \\
\langle st, u \rangle :: \langle \texttt{v}, \texttt{-wh} \rangle_0 &\rightarrow s :: \langle \texttt{=v v} \rangle_1 \quad \langle t, u \rangle :: \langle \texttt{v}, \texttt{-wh} \rangle_0
\end{aligned}$$

Probabilities on MCFGs

$$
\begin{array}{llcll}
\lambda_1 & ts :: \langle \mathtt{c} \rangle_0 & \to & \langle s, t \rangle :: \langle \mathtt{+wh\ c}, \mathtt{-wh} \rangle_0 & \\
\lambda_2 & st :: \langle \mathtt{c} \rangle_0 & \to & s :: \langle \mathtt{=t\ c} \rangle_1 & t :: \langle \mathtt{t} \rangle_0 \\
\lambda_3 & st :: \langle \mathtt{v} \rangle_0 & \to & s :: \langle \mathtt{=d\ v} \rangle_1 & t :: \langle \mathtt{d} \rangle_1 \\
\lambda_4 & st :: \langle \mathtt{v} \rangle_0 & \to & s :: \langle \mathtt{=v\ v} \rangle_1 & t :: \langle \mathtt{v} \rangle_0 \\
\lambda_5 & \langle s, t \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 & \to & s :: \langle \mathtt{=d\ v} \rangle_1 & t :: \langle \mathtt{d\ -wh} \rangle_1 \\
\lambda_6 & \langle st, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 & \to & s :: \langle \mathtt{=v\ v} \rangle_1 & \langle t, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0
\end{array}
$$

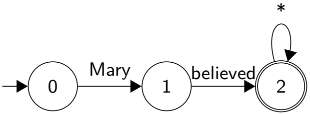
The context-free “backbone” for MG derivations identifies a parametrization for probability distributions over them.

$$
\lambda_2 = \frac{\text{count}\big(\langle \mathtt{c} \rangle_0 \to \langle \mathtt{=t\ c} \rangle_1 \langle \mathtt{t} \rangle_0\big)}{\text{count}\big(\langle \mathtt{c} \rangle_0\big)}
$$

Plus: It turns out that the intersect-with-an-FSA trick we used for CFGs also works for MCFGs!

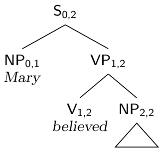
Grammar intersection example (simple)

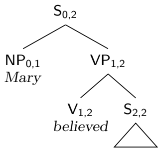
1.0	S → NP VP
0.3	NP → John
0.7	NP → Mary
0.2	VP → ran
0.5	VP → V NP
0.3	VP → V S
0.4	V → believed
0.6	V → knew



1.0	$S_{0,2} \rightarrow NP_{0,1}\ VP_{1,2}$
0.7	$NP_{0,1} \rightarrow$ Mary
0.5	$VP_{1,2} \rightarrow V_{1,2}\ NP_{2,2}$
0.3	$VP_{1,2} \rightarrow V_{1,2}\ S_{2,2}$
0.4	$V_{1,2} \rightarrow$ believed

1.0	$S_{2,2} \rightarrow NP_{2,2}\ VP_{2,2}$
0.3	$NP_{2,2} \rightarrow$ John
0.7	$NP_{2,2} \rightarrow$ Mary
0.2	$VP_{2,2} \rightarrow$ ran
0.5	$VP_{2,2} \rightarrow V_{2,2}\ NP_{2,2}$
0.3	$VP_{2,2} \rightarrow V_{2,2}\ S_{2,2}$
0.4	$V_{2,2} \rightarrow$ believed
0.6	$V_{2,2} \rightarrow$ knew



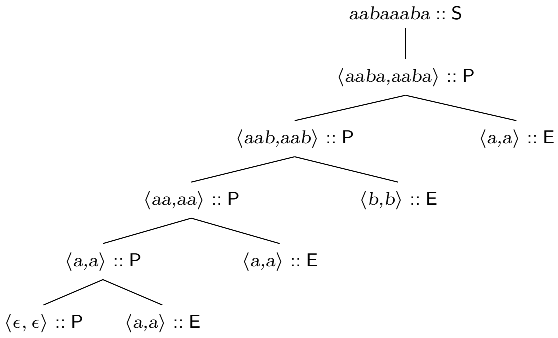


NB: Total weight in this grammar is not one! (What is it? Start symbol is $S_{0,2}$.)
Each derivation has the weight “it” had in the original grammar.

Beyond context-free

$$
\begin{array}{rcl}
t_1 t_2 :: \mathsf{S} & \to & \langle t_1, t_2 \rangle :: \mathsf{P} \\
\langle t_1 u_1, t_2 u_2 \rangle :: \mathsf{P} & \to & \langle t_1, t_2 \rangle :: \mathsf{P} \quad \langle u_1, u_2 \rangle :: \mathsf{E} \\
\langle \epsilon, \epsilon \rangle :: \mathsf{P} & & \\
\langle a, a \rangle :: \mathsf{E} & & \\
\langle b, b \rangle :: \mathsf{E} & &
\end{array}
$$

$$\left\{ ww \mid w \in \{a, b\}^* \right\}$$



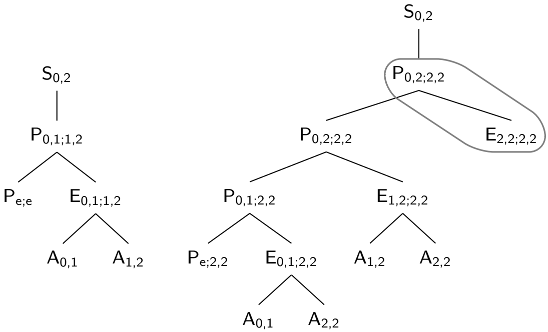
Unlike in a CFG, we can ensure that the two "halves" are extended in the same ways without concatenating them together.

Intersection with an MCFG

$$\begin{array}{lcl} S_{0,2} & \to & P_{0,1;1,2} \\ P_{0,1;1,2} & \to & P_{e;e}\ E_{0,1;1,2} \\ E_{0,1;1,2} & \to & A_{0,1}\ A_{1,2} \end{array}$$

$$\begin{array}{lcl} S_{0,2} & \to & P_{0,2;2,2} \\ P_{0,2;2,2} & \to & P_{0,2;2,2}\ E_{2,2;2,2} \\ P_{0,2;2,2} & \to & P_{0,1;2,2}\ E_{1,2;2,2} \\ P_{0,1;2,2} & \to & P_{e;2,2}\ E_{0,1;2,2} \\ E_{0,1;2,2} & \to & A_{0,1}\ A_{2,2} \\ E_{1,2;2,2} & \to & A_{1,2}\ A_{2,2} \end{array}$$

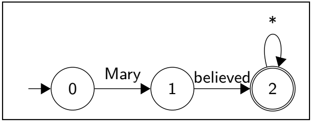
$$\begin{array}{l} \langle b,b \rangle :: E_{2,2;2,2} \\ \langle a,a \rangle :: E_{2,2;2,2} \\ \langle \epsilon, \epsilon \rangle :: P_{e;e} \\ \langle \epsilon, \epsilon \rangle :: P_{e;2,2} \\ a :: A_{2,2} \\ b :: B_{2,2} \\ \\ a :: A_{0,1} \\ a :: A_{1,2} \end{array}$$



Intersection grammars

1.0	S → NP VP
0.3	NP → John
0.7	NP → Mary
0.2	VP → ran
0.5	VP → V NP
0.3	VP → V S
0.4	V → believed
0.6	V → knew

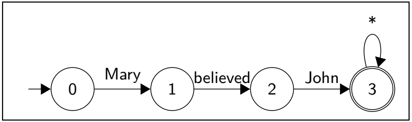
$\cap$



$= \quad G_2$

1.0	S → NP VP
0.3	NP → John
0.7	NP → Mary
0.2	VP → ran
0.5	VP → V NP
0.3	VP → V S
0.4	V → believed
0.6	V → knew

$\cap$



$= \quad G_3$

$$
\begin{aligned}
\text{surprisal at 'John'} &= -\log P(W_3 = \text{John} \mid W_1 = \text{Mary}, W_2 = \text{believed}) \\
&= -\log \frac{\text{total weight in } G_3}{\text{total weight in } G_2} \\
&= -\log \frac{0.0672}{0.224} \\
&= 1.74
\end{aligned}
$$

Surprisal and entropy reduction

$$\begin{aligned}\text{surprisal at `John'} &= -\log P(W_3 = \text{John} \mid W_1 = \text{Mary}, W_2 = \text{believed}) \\ &= -\log \frac{\text{total weight in } G_3}{\text{total weight in } G_2}\end{aligned}$$

$$\text{entropy reduction at `John'} = (\text{entropy of } G_2) - (\text{entropy of } G_3)$$

Computing sum of weights in a grammar ("partition function")

$$Z(A) = \sum_{A \to \alpha} \Big(p(A \to \alpha) \cdot Z(\alpha) \Big)$$
$$Z(\epsilon) = 1$$
$$Z(a\beta) = Z(\beta)$$
$$Z(B\beta) = Z(B) \cdot Z(\beta) \qquad \text{where } \beta \neq \epsilon$$

(Nederhof and Satta 2008)

1.0	S → NP VP
0.3	NP → John
0.7	NP → Mary
0.2	VP → ran
0.5	VP → V NP
0.4	V → believed
0.6	V → knew

$$\begin{aligned} Z(\text{V}) &= 0.4 + 0.6 = 1.0 \\ Z(\text{NP}) &= 0.3 + 0.7 = 1.0 \\ Z(\text{VP}) &= 0.2 + (0.5 \cdot Z(\text{V}) \cdot Z(\text{NP})) \\ &= 0.2 + (0.5 \cdot 1.0 \cdot 1.0) = 0.7 \\ Z(\text{S}) &= 1.0 \cdot Z(\text{NP}) \cdot Z(\text{VP}) \\ &= 0.7 \end{aligned}$$

1.0	S → NP VP
0.3	NP → John
0.7	NP → Mary
0.2	VP → ran
0.5	VP → V NP
0.3	VP → V S
0.4	V → believed
0.6	V → knew

$$\begin{aligned} Z(\text{V}) &= 0.4 + 0.6 = 1.0 \\ Z(\text{NP}) &= 0.3 + 0.7 = 1.0 \\ Z(\text{VP}) &= 0.2 + (0.5 \cdot Z(\text{V}) \cdot Z(\text{NP})) + (0.3 \cdot Z(\text{V}) \cdot Z(\text{S})) \\ Z(\text{S}) &= 1.0 \cdot Z(\text{NP}) \cdot Z(\text{VP}) \end{aligned}$$

Computing entropy of a grammar

1.0	S → NP VP
0.3	NP → John
0.7	NP → Mary
0.2	VP → ran
0.5	VP → V NP
0.3	VP → V S
0.4	V → believed
0.6	V → knew

$$\begin{aligned} h(\mathrm{S}) &= 0 \\ h(\mathrm{NP}) &= \text{entropy of } (0.3, 0.7) \\ h(\mathrm{VP}) &= \text{entropy of } (0.2, 0.5, 0.3) \\ h(\mathrm{V}) &= \text{entropy of } (0.4, 0.6) \end{aligned}$$

$$\begin{aligned} H(\mathrm{S}) &= h(\mathrm{S}) + 1.0(H(\mathrm{NP}) + H(\mathrm{VP})) \\ H(\mathrm{NP}) &= h(\mathrm{NP}) \\ H(\mathrm{VP}) &= h(\mathrm{VP}) + 0.2(0) + 0.5(H(\mathrm{V}) + H(\mathrm{NP})) + 0.3(H(\mathrm{V}) + H(\mathrm{S})) \\ H(\mathrm{V}) &= h(\mathrm{V}) \end{aligned}$$

(Hale 2006)

Surprisal and entropy reduction

$$\begin{aligned}\text{surprisal at `John'} &= -\log P(W_3 = \text{John} \mid W_1 = \text{Mary}, W_2 = \text{believed}) \\ &= -\log \frac{\text{total weight in } G_3}{\text{total weight in } G_2}\end{aligned}$$

$$\text{entropy reduction at `John'} = (\text{entropy of } G_2) - (\text{entropy of } G_3)$$

Putting it all together (Hale 2006)

We can now put entropy reduction/surprisal together with a minimalist grammar to produce predictions about sentence comprehension difficulty!

$$\text{complexity metric} \quad + \quad \text{grammar} \quad \longrightarrow \quad \text{prediction}$$

- Write an MG that generates sentence types of interest
- Convert MG to an MCFG
- Add probabilities to MCFG based on corpus frequencies (or whatever else)
- Compute intersection grammars for each point in a sentence
- Calculate reduction in entropy across the course of the sentence (i.e. workload)

Demo

Hale (2006)

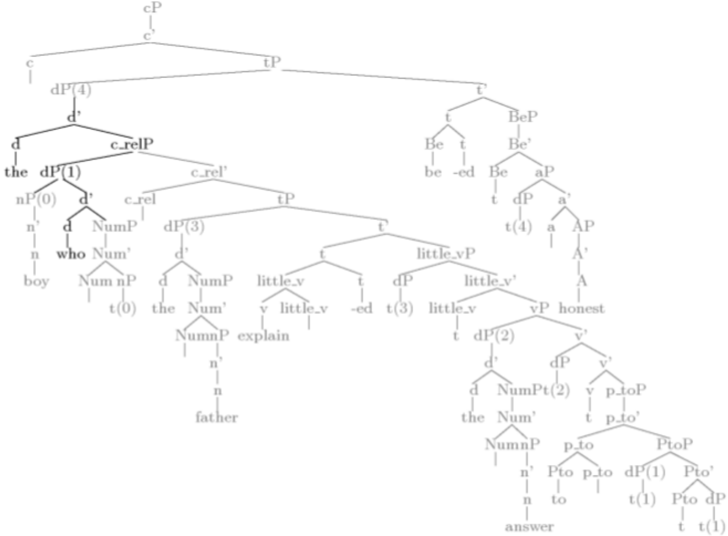


Fig. 11. Kaynian promotion analysis.

Hale (2006)

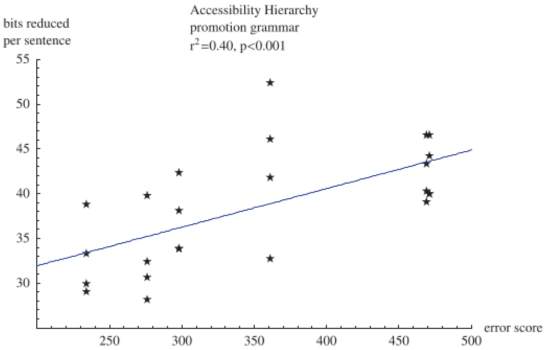
```
they have -ed forget -en that the boy who tell -ed the story be -s so young
the fact that the girl who pay -ed for the ticket be -s very poor doesnt matter
I know that the girl who get -ed the right answer be -s clever
he remember -ed that the man who sell -ed the house leave -ed the town

they have -ed forget -en that the letter which Dick write -ed yesterday be -s long
the fact that the cat which David show -ed to the man like -s eggs be -s strange
I know that the dog which Penny buy -ed today be -s very gentle
he remember -ed that the sweet which David give -ed Sally be -ed a treat

they have -ed forget -en that the man who Ann give -ed the present to be -ed old
the fact that the boy who Paul sell -ed the book to hate -s reading be -s strange
I know that the man who Stephen explain -ed the accident to be -s kind
he remember -ed that the dog which Mary teach -ed the trick to be -s clever

they have -ed forget -en that the box which Pat bring -ed the apple in be -ed lost
the fact that the girl who Sue write -ed the story with be -s proud doesnt matter
I know that the ship which my uncle take -ed Joe on be -ed interesting
he remember -ed that the food which Chris pay -ed the bill for be -ed cheap

they have -ed forget -en that the girl whose friend buy -ed the cake be -ed wait -ing
the fact that the boy whose brother tell -s lies be -s always honest surprise -ed us
I know that the boy whose father sell -ed the dog be -ed very sad
he remember -ed that the girl whose mother send -ed the clothe come -ed too late

they have -ed forget -en that the man whose house Patrick buy -ed be -ed so ill
the fact that the sailor whose ship Jim take -ed have -ed one leg be -s important
I know that the woman whose car Jenny sell -ed be -ed very angry
he remember -ed that the girl whose picture Clare show -ed us be -ed pretty
```

Hale (2006)

count	grammatical relation	definition
1430	subject	co-indexed trace is the first daughter of S
929	direct object	co-indexed trace is immediately following sister of a V-node
167	indirect object	co-indexed trace is part of a PP not annotated as benefactive, locative, manner, purpose, temporal or directional
41	oblique	co-indexed trace is part of a benefactive, locative, manner, purpose, temporal or directional PP
34	genitive subject	WH word is *whose* and co-indexed trace is first daughter of S
4	genitive direct object	WH word is *whose* and co-indexed trace is immediately following sister of a V-node

Fig. 13. Counts from Brown portion of Penn Treebank III.

Hale (2006)



(a) Per-sentence entropy reductions on the promotion grammar

Grammatical Relation:	SU	DO	IO	OBL	GenS	GenO
Repetition Accuracy:	406	364	342	279	167	171
errors (= $R.A._{max}$ − R.A.)	234	276	298	361	471	469

Fig. 8. Results from Keenan and Hawkins (1987).

Hale (2006)

Hale actually wrote two different MGs:

- classical adjunction analysis of relative clauses
- Kaynian/promotion analysis

Hale (2006)

Hale actually wrote two different MGs:

- classical adjunction analysis of relative clauses
- Kaynian/promotion analysis

The branching structure of the two MCFGs was different enough to produce distinct Entropy Reduction predictions. (Same corpus counts!)

The Kaynian/promotion analysis produced a better fit for the Accessibility Hierarchy facts.
(i.e. holding the complexity metric fixed to argue for a grammar)

But there are some ways in which this method is insensitive to fine details of the MG formalism.

Outline

13 Easy probabilities with context-free structure

14 Different frameworks

15 Problem #1 with the naive parametrization

16 Problem #2 with the naive parametrization

17 Solution: Faithfulness to MG operations

Subtlely different minimalist frameworks

Minimalist grammars with many choices of different bells and whistles can all be expressed with context-free derivational structure.

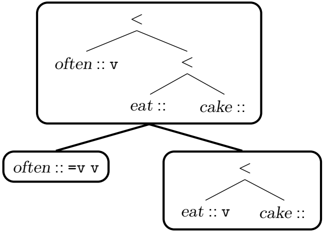
- Must keep an eye on finiteness of number of types (SMC or equivalent)!
- See Stabler (2011)

Some points of variation:

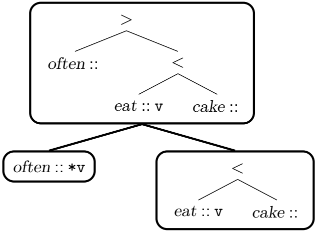
- adjunction
- head movement
- phases
- move as re-merge
- . . .

How to deal with adjuncts?

A normal application of MERGE?

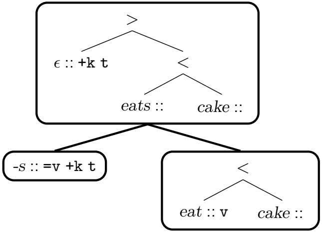


Or a new kind of feature and distinct operation ADJOIN?



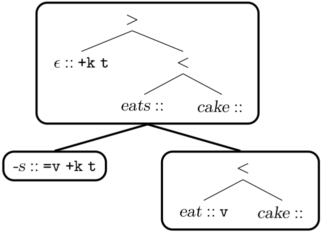
How to implement "head movement"?

Modify MERGE to allow some additional string-shuffling in head-complement relationships?

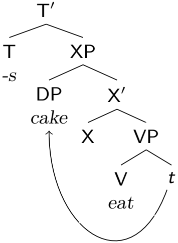


How to implement "head movement"?

Modify MERGE to allow some additional string-shuffling in head-complement relationships?

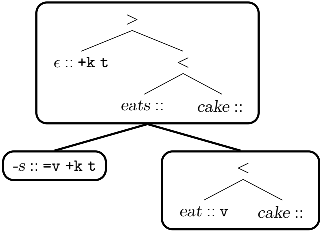


Or some combination of normal phrasal movements? (Koopman and Szabolcsi 2000)

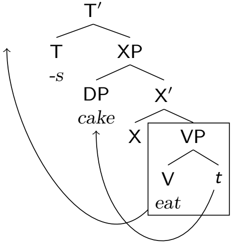


How to implement “head movement”?

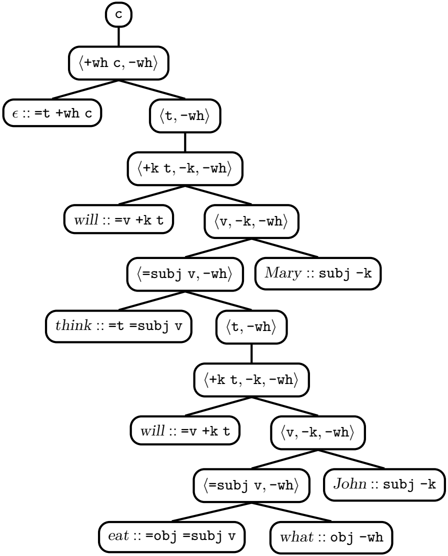
Modify MERGE to allow some additional string-shuffling in head-complement relationships?



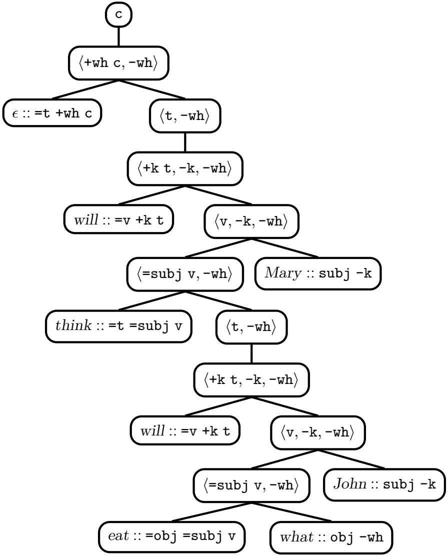
Or some combination of normal phrasal movements? (Koopman and Szabolcsi 2000)

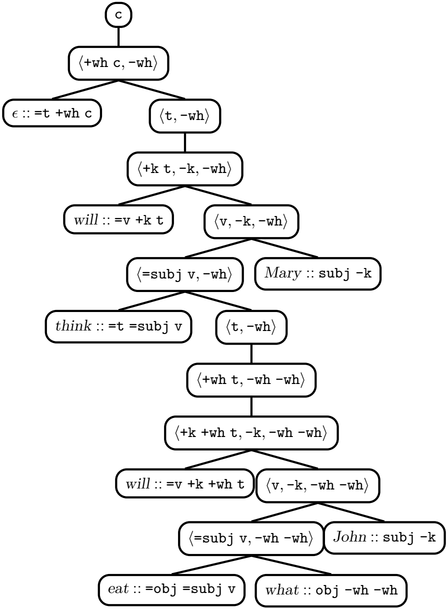


Successive cyclic movement?

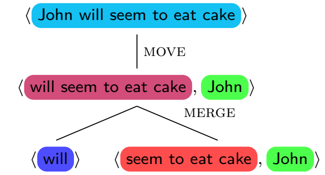


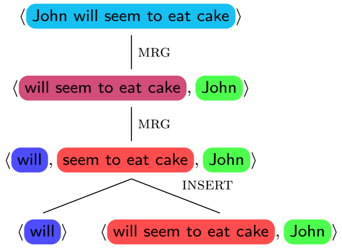
Successive cyclic movement?





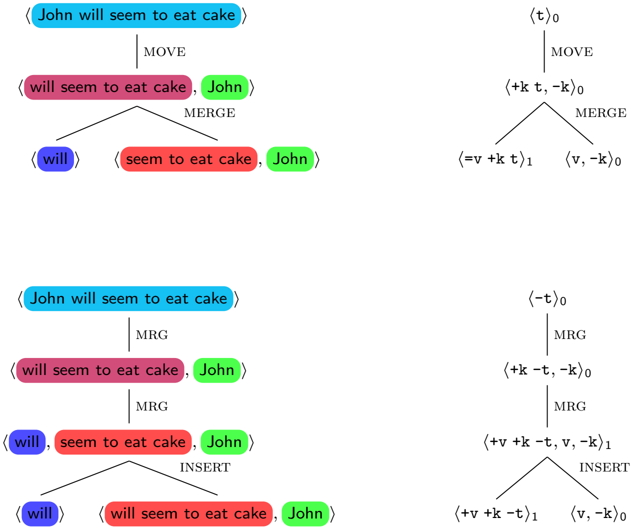
Unifying feature-checking (one way)





(Stabler 2006, Hunter 2011)

Unifying feature-checking (one way)



(Stabler 2006, Hunter 2011)

Three schemas for MERGE rules:

$$\langle st, t_1, \ldots, t_k \rangle :: \langle \gamma, \alpha_1, \ldots, \alpha_k \rangle_0 \quad \rightarrow \quad s :: \langle \texttt{=f}\gamma \rangle_1 \quad \langle t, t_1, \ldots, t_k \rangle :: \langle \texttt{f}, \alpha_1, \ldots, \alpha_k \rangle_n$$

$$\langle ts, s_1, \ldots, s_j, t_1, \ldots, t_k \rangle :: \langle \gamma, \alpha_1, \ldots, \alpha_j, \beta_1, \ldots, \beta_k \rangle_0 \quad \rightarrow \quad \langle s, s_1, \ldots, s_j \rangle :: \langle \texttt{=f}\gamma, \alpha_1, \ldots, \alpha_j \rangle_0 \quad \langle t, t_1, \ldots, t_k \rangle :: \langle \texttt{f}, \beta_1, \ldots, \beta_k \rangle_n$$

$$\langle s, s_1, \ldots, s_j, t, t_1, \ldots, t_k \rangle :: \langle \gamma, \alpha_1, \ldots, \alpha_j, \delta, \beta_1, \ldots, \beta_k \rangle_0 \quad \rightarrow \quad \langle s, s_1, \ldots, s_j \rangle :: \langle \texttt{=f}\gamma, \alpha_1, \ldots, \alpha_j \rangle_n \quad \langle t, t_1, \ldots, t_k \rangle :: \langle \texttt{f}\delta, \beta_1, \ldots, \beta_k \rangle_{n'}$$

Two schemas for MERGE rules:

$$\langle s_i s, s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_k \rangle :: \langle \gamma, \alpha_1, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_k \rangle_0 \quad \rightarrow \quad \langle s, s_1, \ldots, s_i, \ldots, s_k \rangle :: \langle \texttt{+f}\gamma, \alpha_1, \ldots, \alpha_{i-1}, \texttt{-f}, \alpha_{i+1}, \ldots, \alpha_k \rangle_0$$

$$\langle s, s_1, \ldots, s_i, \ldots, s_k \rangle :: \langle \gamma, \alpha_1, \ldots, \alpha_{i-1}, \delta, \alpha_{i+1}, \ldots, \alpha_k \rangle_0 \quad \rightarrow \quad \langle s, s_1, \ldots, s_i, \ldots, s_k \rangle :: \langle \texttt{+f}\gamma, \alpha_1, \ldots, \alpha_{i-1}, \texttt{-f}\delta, \alpha_{i+1}, \ldots, \alpha_k \rangle_0$$

One schema for INSERT rules:

$$\begin{array}{r}\langle s, s_1, \ldots, s_j, t, t_1, \ldots, t_k\rangle :: \langle \texttt{+f}\gamma, \alpha_1, \ldots, \alpha_j, \texttt{-f}\gamma', \beta_1, \ldots, \beta_k\rangle_n \quad \rightarrow \\ s, s_1, \ldots, s_j :: \langle \texttt{+f}\gamma, \alpha_1, \ldots, \alpha_j\rangle_n \quad \langle t, t_1, \ldots, t_k\rangle :: \langle \texttt{-f}\gamma', \beta_1, \ldots, \beta_k\rangle_{n'}\end{array}$$

Three schemas for MRG rules:

$$\begin{array}{r}\langle ss_i, s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_k\rangle :: \langle \gamma, \alpha_1, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_k\rangle_0 \quad \rightarrow \\ \langle s, s_1, \ldots, s_i, \ldots, s_k\rangle :: \langle \texttt{+f}\gamma, \alpha_1, \ldots, \alpha_{i-1}, \texttt{-f}, \alpha_{i+1}, \ldots, \alpha_k\rangle_1\end{array}$$

$$\begin{array}{r}\langle s_i s, s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_k\rangle :: \langle \gamma, \alpha_1, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_k\rangle_0 \quad \rightarrow \\ \langle s, s_1, \ldots, s_i, \ldots, s_k\rangle :: \langle \texttt{+f}\gamma, \alpha_1, \ldots, \alpha_{i-1}, \texttt{-f}, \alpha_{i+1}, \ldots, \alpha_k\rangle_0\end{array}$$

$$\begin{array}{r}\langle s, s_1, \ldots, s_i, \ldots, s_k\rangle :: \langle \gamma, \alpha_1, \ldots, \alpha_{i-1}, \delta, \alpha_{i+1}, \ldots, \alpha_k\rangle_0 \quad \rightarrow \\ \langle s, s_1, \ldots, s_i, \ldots, s_k\rangle :: \langle \texttt{+f}\gamma, \alpha_1, \ldots, \alpha_{i-1}, \texttt{-f}\delta, \alpha_{i+1}, \ldots, \alpha_k\rangle_0\end{array}$$

Subtlely different minimalist frameworks

Minimalist grammars with many choices of different bells and whistles can all be expressed with context-free derivational structure.

- Must keep an eye on finiteness of number of types (SMC or equivalent)!
- See Stabler (2011)

Some points of variation:

- adjunction
- head movement
- phases
- move as re-merge
- ...

Subtlely different minimalist frameworks

Minimalist grammars with many choices of different bells and whistles can all be expressed with context-free derivational structure.

- Must keep an eye on finiteness of number of types (SMC or equivalent)!
- See Stabler (2011)

Some points of variation:

- adjunction
- head movement
- phases
- move as re-merge
- ...

Each variant of the formalism expresses a different hypothesis about the set of primitive grammatical operations. (We are looking for ways to tell these apart!)

- The "shapes" of the derivation trees are generally very similar from one variant to the next.
- But variants will make different classifications of the derivational steps involved, according to which operation is being applied.

Outline

13 Easy probabilities with context-free structure

14 Different frameworks

15 Problem #1 with the naive parametrization

16 Problem #2 with the naive parametrization

17 Solution: Faithfulness to MG operations

Probabilities on MCFGs

$$
\begin{array}{lrcl}
\lambda_1 & ts :: \langle \texttt{c} \rangle_0 & \to & \langle s, t \rangle :: \langle \texttt{+wh c}, \texttt{-wh} \rangle_0 \\
\lambda_2 & st :: \langle \texttt{c} \rangle_0 & \to & s :: \langle \texttt{=t c} \rangle_1 \quad t :: \langle \texttt{t} \rangle_0 \\
\lambda_3 & st :: \langle \texttt{v} \rangle_0 & \to & s :: \langle \texttt{=d v} \rangle_1 \quad t :: \langle \texttt{d} \rangle_1 \\
\lambda_4 & st :: \langle \texttt{v} \rangle_0 & \to & s :: \langle \texttt{=v v} \rangle_1 \quad t :: \langle \texttt{v} \rangle_0 \\
\lambda_5 & \langle s, t \rangle :: \langle \texttt{v}, \texttt{-wh} \rangle_0 & \to & s :: \langle \texttt{=d v} \rangle_1 \quad t :: \langle \texttt{d -wh} \rangle_1 \\
\lambda_6 & \langle st, u \rangle :: \langle \texttt{v}, \texttt{-wh} \rangle_0 & \to & s :: \langle \texttt{=v v} \rangle_1 \quad \langle t, u \rangle :: \langle \texttt{v}, \texttt{-wh} \rangle_0
\end{array}
$$

Training question: What values of λ_1, λ_2, etc. make the training corpus most likely?

Problem #1 with the naive parametrization

Grammar

pierre :: `d`	*who* :: `d -wh`
marie :: `d`	*will* :: `=v =d t`
praise :: `=d v`	ϵ :: `=t c`
often :: `=v v`	ϵ :: `=t +wh c`

Training data

90	pierre will praise marie
5	pierre will **often** praise marie
1	who pierre will praise
1	who pierre will **often** praise

Problem #1 with the naive parametrization

Grammar

pierre :: d	*who* :: d -wh
marie :: d	*will* :: =v =d t
praise :: =d v	ϵ :: =t c
often :: =v v	ϵ :: =t +wh c

Training data

90	pierre will praise marie
5	pierre will **often** praise marie
1	who pierre will praise
1	who pierre will **often** praise

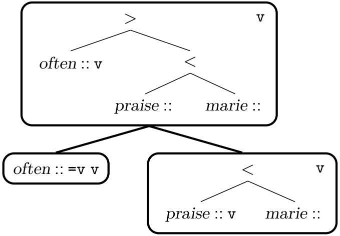
$$st :: \langle \mathtt{v} \rangle_0 \rightarrow s :: \langle \mathtt{=d\ v} \rangle_1 \quad t :: \langle \mathtt{d} \rangle_1 \qquad 0.95$$

$$st :: \langle \mathtt{v} \rangle_0 \rightarrow s :: \langle \mathtt{=v\ v} \rangle_1 \quad t :: \langle \mathtt{v} \rangle_0 \qquad 0.05$$

$$\langle s, t \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 \rightarrow s :: \langle \mathtt{=d\ v} \rangle_1 \quad t :: \langle \mathtt{d\ -wh} \rangle_1 \qquad 0.67$$

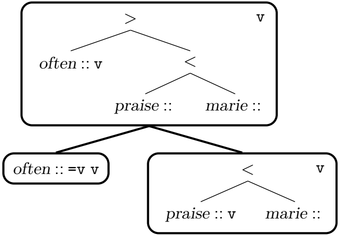
$$\langle st, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 \rightarrow s :: \langle \mathtt{=v\ v} \rangle_1 \quad \langle t, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 \qquad 0.33$$

Generalizations missed by the naive parametrization

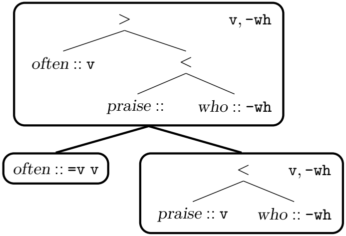


$$\textit{st} :: \langle \mathtt{v} \rangle_0 \quad \rightarrow \quad \textit{s} :: \langle \mathtt{=v\ v} \rangle_1 \quad \textit{t} :: \langle \mathtt{v} \rangle_0$$

Generalizations missed by the naive parametrization



$$\mathbf{st} :: \langle \mathtt{v} \rangle_0 \quad \rightarrow \quad \mathbf{s} :: \langle \mathtt{=v\ v} \rangle_1 \quad \mathbf{t} :: \langle \mathtt{v} \rangle_0$$



$$\langle \mathbf{st}, \mathbf{u} \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 \quad \rightarrow \quad \mathbf{s} :: \langle \mathtt{=v\ v} \rangle_1 \quad \langle \mathbf{t}, \mathbf{u} \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0$$

Problem #1 with the naive parametrization

Grammar

pierre :: d	*who* :: d -wh
marie :: d	*will* :: =v =d t
praise :: =d v	ϵ :: =t c
often :: =v v	ϵ :: =t +wh c

Training data

90	pierre will praise marie
5	pierre will **often** praise marie
1	who pierre will praise
1	who pierre will **often** praise

$$
\begin{array}{rcllr}
st :: \langle \mathtt{v} \rangle_0 & \rightarrow & s :: \langle \mathtt{=d\ v} \rangle_1 & t :: \langle \mathtt{d} \rangle_1 & 0.95 \\
st :: \langle \mathtt{v} \rangle_0 & \rightarrow & s :: \langle \mathtt{=v\ v} \rangle_1 & t :: \langle \mathtt{v} \rangle_0 & 0.05 \\
\langle s, t \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 & \rightarrow & s :: \langle \mathtt{=d\ v} \rangle_1 & t :: \langle \mathtt{d\ -wh} \rangle_1 & 0.67 \\
\langle st, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 & \rightarrow & s :: \langle \mathtt{=v\ v} \rangle_1 & \langle t, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 & 0.33
\end{array}
$$

Problem #1 with the naive parametrization

Grammar

pierre :: `d`	*who* :: `d -wh`
marie :: `d`	*will* :: `=v =d t`
praise :: `=d v`	ϵ :: `=t c`
often :: `=v v`	ϵ :: `=t +wh c`

Training data

90	pierre will praise marie
5	pierre will **often** praise marie
1	who pierre will praise
1	who pierre will **often** praise

$$\begin{array}{rcll} st :: \langle \mathtt{v} \rangle_0 & \rightarrow & s :: \langle \mathtt{=d\ v} \rangle_1 \quad t :: \langle \mathtt{d} \rangle_1 & 0.95 \\ st :: \langle \mathtt{v} \rangle_0 & \rightarrow & s :: \langle \mathtt{=v\ v} \rangle_1 \quad t :: \langle \mathtt{v} \rangle_0 & 0.05 \\ \langle s, t \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 & \rightarrow & s :: \langle \mathtt{=d\ v} \rangle_1 \quad t :: \langle \mathtt{d\ -wh} \rangle_1 & 0.67 \\ \langle st, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 & \rightarrow & s :: \langle \mathtt{=v\ v} \rangle_1 \quad \langle t, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 & 0.33 \end{array}$$

$$\frac{\text{count}\Big(\langle \mathtt{v} \rangle_0 \rightarrow \langle \mathtt{=d\ v} \rangle_1\ \langle \mathtt{d} \rangle_1\Big)}{\text{count}\Big(\langle \mathtt{v} \rangle_0\Big)} = \frac{95}{100}$$

$$\frac{\text{count}\Big(\langle \mathtt{v,-wh} \rangle_0 \rightarrow \langle \mathtt{=d\ v} \rangle_1\ \langle \mathtt{d\ -wh} \rangle_1\Big)}{\text{count}\Big(\langle \mathtt{v,-wh} \rangle_0\Big)} = \frac{2}{3}$$

This training setup doesn't know which minimalist-grammar operations are being implemented by the various MCFG rules.

Naive parametrization

$$\begin{tikzcd}
G_A \arrow[r] & \begin{array}{c}\text{Naive}\\\text{parametrization}\end{array} \arrow[r] & \begin{array}{l}0.95\\0.05\\0.67\\0.33\end{array} \\
& \text{Training corpus} \arrow[u] &
\end{tikzcd}$$

Outline

13 Easy probabilities with context-free structure

14 Different frameworks

15 Problem #1 with the naive parametrization

16 Problem #2 with the naive parametrization

17 Solution: Faithfulness to MG operations

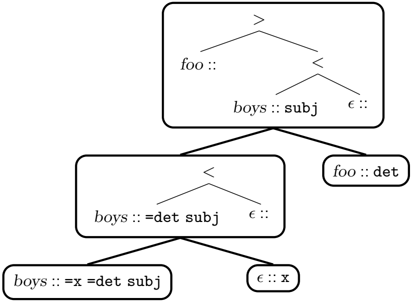
A (slightly) more complicated grammar

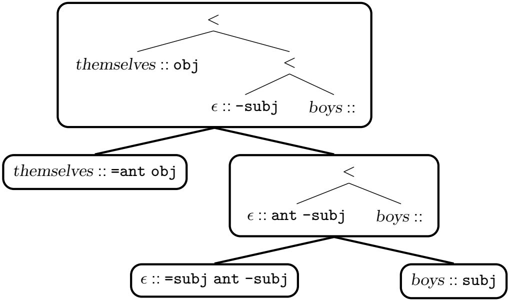
ϵ :: `=t c`	*boys* :: `=x =det subj`
ϵ :: `=t +wh c`	ϵ :: `x`
will :: `=v =subj t`	*foo* :: `det`
shave :: `v`	
shave :: `=obj v`	*themselves* :: `=ant obj`
boys :: `subj`	ϵ :: `=subj ant -subj`
who :: `subj -wh`	*will* :: `=v +subj t`

boys will shave
boys will shave themselves
who will shave
who will shave themselves
foo boys will shave
foo boys will shave themselves

Some details:

- Subject is base-generated in SpecTP; no movement for Case
- Transitive and intransitive versions of *shave*
- *foo* is a determiner that optionally combines with *boys* to make a subject
  - Dummy feature `x` to fill complement of *boys* so that *foo* goes on the left
- *themselves* can appear in object position, via a movement theory of reflexives
  - A `subj` can be turned into an `ant -subj`
  - *themselves* combines with an `ant` to make an `obj`
  - *will* can attract its subject by move as well as merge


>
foo ::
<
boys :: subj
ϵ ::
<
boys :: =det subj
ϵ ::
foo :: det
boys :: =x =det subj
ϵ :: x


<
themselves :: obj
<
ϵ :: -subj
boys ::
themselves :: =ant obj
<
ϵ :: ant -subj
boys ::
ϵ :: =subj ant -subj
boys :: subj

Choice points in the MG-derived MCFG

Question or not?			
$\langle \texttt{c} \rangle_0$	$\rightarrow$	$\langle \texttt{=t c} \rangle_0$	$\langle \texttt{t} \rangle_0$
$\langle \texttt{c} \rangle_0$	$\rightarrow$	$\langle \texttt{+wh c}, \texttt{-wh} \rangle_0$	

Antecedent lexical or complex?			
$\langle \texttt{ant -subj} \rangle_0$	$\rightarrow$	$\langle \texttt{=subj ant -subj} \rangle_1$	$\langle \texttt{subj} \rangle_0$
$\langle \texttt{ant -subj} \rangle_0$	$\rightarrow$	$\langle \texttt{=subj ant -subj} \rangle_1$	$\langle \texttt{subj} \rangle_1$

Non-wh subject merged and complex, merged and lexical, or moved?			
$\langle \texttt{t} \rangle_0$	$\rightarrow$	$\langle \texttt{=subj t} \rangle_0$	$\langle \texttt{subj} \rangle_0$
$\langle \texttt{t} \rangle_0$	$\rightarrow$	$\langle \texttt{=subj t} \rangle_0$	$\langle \texttt{subj} \rangle_1$
$\langle \texttt{t} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj t}, \texttt{-subj} \rangle_0$	

Wh-phrase same as moving subject or separated because of doubling?			
$\langle \texttt{t}, \texttt{-wh} \rangle_0$	$\rightarrow$	$\langle \texttt{=subj t} \rangle_0$	$\langle \texttt{subj -wh} \rangle_1$
$\langle \texttt{t}, \texttt{-wh} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj t}, \texttt{-subj}, \texttt{-wh} \rangle_0$	

Choice points in the IMG-derived MCFG

Question or not?		
$\langle \texttt{-c} \rangle_0$	$\rightarrow$	$\langle \texttt{+t -c}, \texttt{-t} \rangle_1$
$\langle \texttt{-c} \rangle_0$	$\rightarrow$	$\langle \texttt{+wh -c}, \texttt{-wh} \rangle_0$

Antecedent lexical or complex?			
$\langle \texttt{+subj -ant -subj}, \texttt{-subj} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj -ant -subj} \rangle_0$	$\langle \texttt{-subj} \rangle_0$
$\langle \texttt{+subj -ant -subj}, \texttt{-subj} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj -ant -subj} \rangle_0$	$\langle \texttt{-subj} \rangle_1$

Non-wh subject merged and complex, merged and lexical, or moved?			
$\langle \texttt{+subj -t}, \texttt{-subj} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj -t} \rangle_0$	$\langle \texttt{-subj} \rangle_0$
$\langle \texttt{+subj -t}, \texttt{-subj} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj -t} \rangle_0$	$\langle \texttt{-subj} \rangle_1$
$\langle \texttt{+subj -t}, \texttt{-subj} \rangle_0$	$\rightarrow$	$\langle \texttt{+v +subj -t}, \texttt{-v}, \texttt{-subj} \rangle_1$	

Wh-phrase same as moving subject or separated because of doubling?		
$\langle \texttt{-t}, \texttt{-wh} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj -t}, \texttt{-subj -wh} \rangle_0$
$\langle \texttt{-t}, \texttt{-wh} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj -t}, \texttt{-subj}, \texttt{-wh} \rangle_0$

Problem #2 with the naive parametrization

$$
\begin{tikzcd}
\boxed{\text{``normal'' MG}} \arrow[r] \arrow[d] & \boxed{\text{MCFG}} \\
\boxed{\text{``re-merge'' MG}} \arrow[r] & \boxed{\text{MCFG}}
\end{tikzcd}
$$

Language of both grammars

boys will shave
boys will shave themselves
who will shave
who will shave themselves
foo boys will shave
foo boys will shave themselves

Training data

10	boys will shave
2	boys will shave themselves
3	who will shave
1	who will shave themselves
5	foo boys will shave

Problem #2 with the naive parametrization

$$\begin{tikzcd} \boxed{\text{``normal'' MG}} \arrow[r] \arrow[d] & \boxed{\text{MCFG}} \\ \boxed{\text{``re-merge'' MG}} \arrow[r] & \boxed{\text{MCFG}} \end{tikzcd}$$

Language of both grammars

boys will shave
boys will shave themselves
who will shave
who will shave themselves
foo boys will shave
foo boys will shave themselves

Training data

10	boys will shave
2	boys will shave themselves
3	who will shave
1	who will shave themselves
5	foo boys will shave

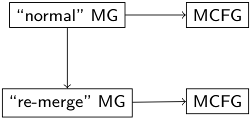
With merge and move distinct operations:

0.47619	boys will shave
0.238095	foo boys will shave
0.142857	who will shave
0.0952381	boys will shave themselves
0.047619	who will shave themselves

With merge and move as unified operations:

0.47619	boys will shave
0.238095	foo boys will shave
0.142857	who will shave
0.0952381	boys will shave themselves
0.047619	who will shave themselves

Problem #2 with the naive parametrization



Language of both grammars

boys will shave
boys will shave themselves
who will shave
who will shave themselves
foo boys will shave
foo boys will shave themselves

Training data

10	boys will shave
2	boys will shave themselves
3	who will shave
1	who will shave themselves
5	foo boys will shave

With merge and move distinct operations:

0.47619	boys will shave
0.238095	foo boys will shave
0.142857	who will shave
0.0952381	boys will shave themselves
0.047619	who will shave themselves

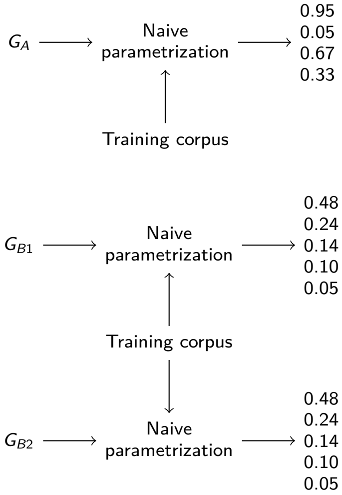
With merge and move as unified operations:

0.47619	boys will shave
0.238095	foo boys will shave
0.142857	who will shave
0.0952381	boys will shave themselves
0.047619	who will shave themselves

This treatment of probabilities doesn't know which minimalist-grammar operations are being implemented by the various MCFG rules.

So the probabilities are unaffected by changes in set of primitive operations.

Naive parametrization



Outline

13 Easy probabilities with context-free structure

14 Different frameworks

15 Problem #1 with the naive parametrization

16 Problem #2 with the naive parametrization

17 Solution: Faithfulness to MG operations

The smarter parametrization

Solution: Have a rule's probability be a function of (only) "what it does"

- merge or move
- what feature is being checked (either movement or selection)

(Hunter and Dyer 2013)

The smarter parametrization

Solution: Have a rule's probability be a function of (only) "what it does"

- merge or move
- what feature is being checked (either movement or selection)

MCFG Rule	ϕ_{MERGE}	$\phi_{\mathtt{d}}$	$\phi_{\mathtt{v}}$	$\phi_{\mathtt{t}}$	ϕ_{MOVE}	$\phi_{\mathtt{wh}}$
$st :: \langle \mathtt{c} \rangle_0 \rightarrow s :: \langle \mathtt{=t\ c} \rangle_1 \quad t :: \langle \mathtt{t} \rangle_0$	1	0	0	1	0	0
$ts :: \langle \mathtt{c} \rangle_0 \rightarrow \langle s, t \rangle :: \langle \mathtt{+wh\ c}, \mathtt{-wh} \rangle_0$	0	0	0	0	1	1
$st :: \langle \mathtt{v} \rangle_0 \rightarrow s :: \langle \mathtt{=d\ v} \rangle_1 \quad t :: \langle \mathtt{d} \rangle_1$	1	1	0	0	0	0
$st :: \langle \mathtt{v} \rangle_0 \rightarrow s :: \langle \mathtt{=v\ v} \rangle_1 \quad t :: \langle \mathtt{v} \rangle_0$	1	0	1	0	0	0
$\langle s, t \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 \rightarrow s :: \langle \mathtt{=d\ v} \rangle_1 \quad t :: \langle \mathtt{d\ -wh} \rangle_1$	1	1	0	0	0	0
$\langle st, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 \rightarrow s :: \langle \mathtt{=v\ v} \rangle_1 \quad \langle t, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0$	1	0	1	0	0	0

Each rule r is assigned a score as a function of the vector $\phi(r)$:

$$\begin{aligned} s(r) &= \exp(\boldsymbol{\lambda} \cdot \phi(r)) \\ &= \exp(\lambda_{\text{MERGE}}\, \phi_{\text{MERGE}}(r) + \lambda_{\mathtt{d}}\, \phi_{\mathtt{d}}(r) + \lambda_{\mathtt{v}}\, \phi_{\mathtt{v}}(r) + \dots) \end{aligned}$$

(Hunter and Dyer 2013)

The smarter parametrization

Solution: Have a rule's probability be a function of (only) "what it does"

- merge or move
- what feature is being checked (either movement or selection)

MCFG Rule	ϕ_{MERGE}	$\phi_{\mathtt{d}}$	$\phi_{\mathtt{v}}$	$\phi_{\mathtt{t}}$	ϕ_{MOVE}	$\phi_{\mathtt{wh}}$
$st :: \langle \mathtt{c} \rangle_0 \rightarrow s :: \langle \mathtt{=t\ c} \rangle_1 \quad t :: \langle \mathtt{t} \rangle_0$	1	0	0	1	0	0
$ts :: \langle \mathtt{c} \rangle_0 \rightarrow \langle s, t \rangle :: \langle \mathtt{+wh\ c}, \mathtt{-wh} \rangle_0$	0	0	0	0	1	1
$st :: \langle \mathtt{v} \rangle_0 \rightarrow s :: \langle \mathtt{=d\ v} \rangle_1 \quad t :: \langle \mathtt{d} \rangle_1$	1	1	0	0	0	0
$st :: \langle \mathtt{v} \rangle_0 \rightarrow s :: \langle \mathtt{=v\ v} \rangle_1 \quad t :: \langle \mathtt{v} \rangle_0$	1	0	1	0	0	0
$\langle s, t \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 \rightarrow s :: \langle \mathtt{=d\ v} \rangle_1 \quad t :: \langle \mathtt{d\ -wh} \rangle_1$	1	1	0	0	0	0
$\langle st, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 \rightarrow s :: \langle \mathtt{=v\ v} \rangle_1 \quad \langle t, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0$	1	0	1	0	0	0

Each rule r is assigned a score as a function of the vector $\phi(r)$:

$$s(r) = \exp(\boldsymbol{\lambda} \cdot \phi(r))$$

$$= \exp(\lambda_{\text{MERGE}}\, \phi_{\text{MERGE}}(r) + \lambda_{\mathtt{d}}\, \phi_{\mathtt{d}}(r) + \lambda_{\mathtt{v}}\, \phi_{\mathtt{v}}(r) + \dots)$$

$$s(r_1) = \exp(\lambda_{\text{MERGE}} + \lambda_{\mathtt{t}})$$

(Hunter and Dyer 2013)

The smarter parametrization

Solution: Have a rule's probability be a function of (only) "what it does"

- merge or move
- what feature is being checked (either movement or selection)

MCFG Rule	ϕ_{MERGE}	$\phi_{\mathtt{d}}$	$\phi_{\mathtt{v}}$	$\phi_{\mathtt{t}}$	ϕ_{MOVE}	$\phi_{\mathtt{wh}}$
$st :: \langle \mathtt{c} \rangle_0 \rightarrow s :: \langle \mathtt{=t\ c} \rangle_1 \quad t :: \langle \mathtt{t} \rangle_0$	1	0	0	1	0	0
$ts :: \langle \mathtt{c} \rangle_0 \rightarrow \langle s, t \rangle :: \langle \mathtt{+wh\ c}, \mathtt{-wh} \rangle_0$	0	0	0	0	1	1
$st :: \langle \mathtt{v} \rangle_0 \rightarrow s :: \langle \mathtt{=d\ v} \rangle_1 \quad t :: \langle \mathtt{d} \rangle_1$	1	1	0	0	0	0
$st :: \langle \mathtt{v} \rangle_0 \rightarrow s :: \langle \mathtt{=v\ v} \rangle_1 \quad t :: \langle \mathtt{v} \rangle_0$	1	0	1	0	0	0
$\langle s, t \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 \rightarrow s :: \langle \mathtt{=d\ v} \rangle_1 \quad t :: \langle \mathtt{d\ -wh} \rangle_1$	1	1	0	0	0	0
$\langle st, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 \rightarrow s :: \langle \mathtt{=v\ v} \rangle_1 \quad \langle t, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0$	1	0	1	0	0	0

Each rule r is assigned a score as a function of the vector $\phi(r)$:

$$s(r) = \exp(\boldsymbol{\lambda} \cdot \phi(r))$$
$$= \exp(\lambda_{\text{MERGE}}\, \phi_{\text{MERGE}}(r) + \lambda_{\mathtt{d}}\, \phi_{\mathtt{d}}(r) + \lambda_{\mathtt{v}}\, \phi_{\mathtt{v}}(r) + \dots)$$

$$s(r_1) = \exp(\lambda_{\text{MERGE}} + \lambda_{\mathtt{t}})$$
$$s(r_2) = \exp(\lambda_{\text{MOVE}} + \lambda_{\mathtt{wh}})$$

(Hunter and Dyer 2013)

The smarter parametrization

Solution: Have a rule's probability be a function of (only) "what it does"

- merge or move
- what feature is being checked (either movement or selection)

MCFG Rule	ϕ_{MERGE}	$\phi_{\mathtt{d}}$	$\phi_{\mathtt{v}}$	$\phi_{\mathtt{t}}$	ϕ_{MOVE}	$\phi_{\mathtt{wh}}$
$st :: \langle \mathtt{c} \rangle_0 \rightarrow s :: \langle \mathtt{=t\ c} \rangle_1 \quad t :: \langle \mathtt{t} \rangle_0$	1	0	0	1	0	0
$ts :: \langle \mathtt{c} \rangle_0 \rightarrow \langle s, t \rangle :: \langle \mathtt{+wh\ c}, \mathtt{-wh} \rangle_0$	0	0	0	0	1	1
$st :: \langle \mathtt{v} \rangle_0 \rightarrow s :: \langle \mathtt{=d\ v} \rangle_1 \quad t :: \langle \mathtt{d} \rangle_1$	1	1	0	0	0	0
$st :: \langle \mathtt{v} \rangle_0 \rightarrow s :: \langle \mathtt{=v\ v} \rangle_1 \quad t :: \langle \mathtt{v} \rangle_0$	1	0	1	0	0	0
$\langle s, t \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 \rightarrow s :: \langle \mathtt{=d\ v} \rangle_1 \quad t :: \langle \mathtt{d\ -wh} \rangle_1$	1	1	0	0	0	0
$\langle st, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 \rightarrow s :: \langle \mathtt{=v\ v} \rangle_1 \quad \langle t, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0$	1	0	1	0	0	0

Each rule r is assigned a score as a function of the vector $\phi(r)$:

$$s(r) = \exp(\boldsymbol{\lambda} \cdot \phi(r))$$

$$= \exp(\lambda_{\text{MERGE}}\, \phi_{\text{MERGE}}(r) + \lambda_{\mathtt{d}}\, \phi_{\mathtt{d}}(r) + \lambda_{\mathtt{v}}\, \phi_{\mathtt{v}}(r) + \dots)$$

$$s(r_1) = \exp(\lambda_{\text{MERGE}} + \lambda_{\mathtt{t}})$$

$$s(r_2) = \exp(\lambda_{\text{MOVE}} + \lambda_{\mathtt{wh}})$$

$$s(r_3) = \exp(\lambda_{\text{MERGE}} + \lambda_{\mathtt{d}})$$

(Hunter and Dyer 2013)

The smarter parametrization

Solution: Have a rule's probability be a function of (only) "what it does"

- merge or move
- what feature is being checked (either movement or selection)

MCFG Rule	ϕ_{MERGE}	$\phi_{\texttt{d}}$	$\phi_{\texttt{v}}$	$\phi_{\texttt{t}}$	ϕ_{MOVE}	$\phi_{\texttt{wh}}$
$st :: \langle \texttt{c} \rangle_0 \rightarrow s :: \langle \texttt{=t c} \rangle_1 \quad t :: \langle \texttt{t} \rangle_0$	1	0	0	1	0	0
$ts :: \langle \texttt{c} \rangle_0 \rightarrow \langle s, t \rangle :: \langle \texttt{+wh c}, \texttt{-wh} \rangle_0$	0	0	0	0	1	1
$st :: \langle \texttt{v} \rangle_0 \rightarrow s :: \langle \texttt{=d v} \rangle_1 \quad t :: \langle \texttt{d} \rangle_1$	1	1	0	0	0	0
$st :: \langle \texttt{v} \rangle_0 \rightarrow s :: \langle \texttt{=v v} \rangle_1 \quad t :: \langle \texttt{v} \rangle_0$	1	0	1	0	0	0
$\langle s, t \rangle :: \langle \texttt{v}, \texttt{-wh} \rangle_0 \rightarrow s :: \langle \texttt{=d v} \rangle_1 \quad t :: \langle \texttt{d -wh} \rangle_1$	1	1	0	0	0	0
$\langle st, u \rangle :: \langle \texttt{v}, \texttt{-wh} \rangle_0 \rightarrow s :: \langle \texttt{=v v} \rangle_1 \quad \langle t, u \rangle :: \langle \texttt{v}, \texttt{-wh} \rangle_0$	1	0	1	0	0	0

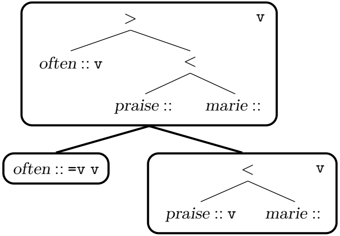
Each rule r is assigned a score as a function of the vector $\phi(r)$:

$$\begin{aligned} s(r) &= \exp(\boldsymbol{\lambda} \cdot \phi(r)) \\ &= \exp(\lambda_{\text{MERGE}}\, \phi_{\text{MERGE}}(r) + \lambda_{\texttt{d}}\, \phi_{\texttt{d}}(r) + \lambda_{\texttt{v}}\, \phi_{\texttt{v}}(r) + \dots) \end{aligned}$$

$$\begin{aligned} s(r_1) &= \exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{t}}) \\ s(r_2) &= \exp(\lambda_{\text{MOVE}} + \lambda_{\texttt{wh}}) \\ s(r_3) &= \exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{d}}) \\ s(r_5) &= \exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{d}}) \end{aligned}$$

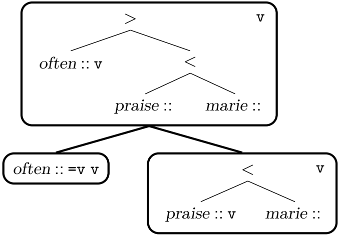
(Hunter and Dyer 2013)

Generalizations missed by the naive parametrization

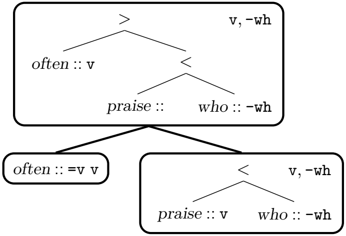


$$\textit{st} :: \langle \mathtt{v} \rangle_0 \quad \longrightarrow \quad \textit{s} :: \langle \mathtt{=v\ v} \rangle_1 \quad \textit{t} :: \langle \mathtt{v} \rangle_0$$

Generalizations missed by the naive parametrization



$$\boldsymbol{st} :: \langle \mathtt{v} \rangle_0 \quad \rightarrow \quad \boldsymbol{s} :: \langle \mathtt{=v\ v} \rangle_1 \quad \boldsymbol{t} :: \langle \mathtt{v} \rangle_0$$



$$\langle \boldsymbol{st}, \boldsymbol{u} \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 \quad \rightarrow \quad \boldsymbol{s} :: \langle \mathtt{=v\ v} \rangle_1 \quad \langle \boldsymbol{t}, \boldsymbol{u} \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0$$

Comparison

The old way:

λ_1	$ts :: \langle \mathtt{c} \rangle_0$	$\rightarrow$	$\langle s, t \rangle :: \langle \mathtt{+wh\ c}, \mathtt{-wh} \rangle_0$
λ_2	$st :: \langle \mathtt{c} \rangle_0$	$\rightarrow$	$s :: \langle \mathtt{=t\ c} \rangle_1 \quad t :: \langle \mathtt{t} \rangle_0$
λ_3	$st :: \langle \mathtt{v} \rangle_0$	$\rightarrow$	$s :: \langle \mathtt{=d\ v} \rangle_1 \quad t :: \langle \mathtt{d} \rangle_1$
λ_4	$st :: \langle \mathtt{v} \rangle_0$	$\rightarrow$	$s :: \langle \mathtt{=v\ v} \rangle_1 \quad t :: \langle \mathtt{v} \rangle_0$
λ_5	$\langle s, t \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0$	$\rightarrow$	$s :: \langle \mathtt{=d\ v} \rangle_1 \quad t :: \langle \mathtt{d\ -wh} \rangle_1$
λ_6	$\langle st, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0$	$\rightarrow$	$s :: \langle \mathtt{=v\ v} \rangle_1 \quad \langle t, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0$

Training question: What values of λ_1, λ_2, etc. make the training corpus most likely?

The new way:

$\exp(\lambda_{\text{MOVE}} + \lambda_{\mathtt{wh}})$	$ts :: \langle \mathtt{c} \rangle_0$	$\rightarrow$	$\langle s, t \rangle :: \langle \mathtt{+wh\ c}, \mathtt{-wh} \rangle_0$
$\exp(\lambda_{\text{MERGE}} + \lambda_{\mathtt{t}})$	$st :: \langle \mathtt{c} \rangle_0$	$\rightarrow$	$s :: \langle \mathtt{=t\ c} \rangle_1 \quad t :: \langle \mathtt{t} \rangle_0$
$\exp(\lambda_{\text{MERGE}} + \lambda_{\mathtt{d}})$	$st :: \langle \mathtt{v} \rangle_0$	$\rightarrow$	$s :: \langle \mathtt{=d\ v} \rangle_1 \quad t :: \langle \mathtt{d} \rangle_1$
$\exp(\lambda_{\text{MERGE}} + \lambda_{\mathtt{v}})$	$st :: \langle \mathtt{v} \rangle_0$	$\rightarrow$	$s :: \langle \mathtt{=v\ v} \rangle_1 \quad t :: \langle \mathtt{v} \rangle_0$
$\exp(\lambda_{\text{MERGE}} + \lambda_{\mathtt{d}})$	$\langle s, t \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0$	$\rightarrow$	$s :: \langle \mathtt{=d\ v} \rangle_1 \quad t :: \langle \mathtt{d\ -wh} \rangle_1$
$\exp(\lambda_{\text{MERGE}} + \lambda_{\mathtt{v}})$	$\langle st, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0$	$\rightarrow$	$s :: \langle \mathtt{=v\ v} \rangle_1 \quad \langle t, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0$

Training question: What values of λ_{MERGE}, λ_{MOVE}, $\lambda_{\mathtt{d}}$, etc. make the training corpus most likely?

Solution #1 with the smarter parametrization

Grammar

pierre :: `d`	*who* :: `d -wh`
marie :: `d`	*will* :: `=v =d t`
praise :: `=d v`	ϵ :: `=t c`
often :: `=v v`	ϵ :: `=t +wh c`

Training data

90	pierre will praise marie
5	pierre will **often** praise marie
1	who pierre will praise
1	who pierre will **often** praise

Maximise likelihood via stochastic gradient ascent:

$$P_{\boldsymbol{\lambda}}(N \to \delta) = \frac{\exp(\boldsymbol{\lambda} \cdot \phi(N \to \delta))}{\sum \exp(\boldsymbol{\lambda} \cdot \phi(N \to \delta'))}$$

Solution #1 with the smarter parametrization

Grammar

pierre :: `d`	*who* :: `d -wh`
marie :: `d`	*will* :: `=v =d t`
praise :: `=d v`	ϵ :: `=t c`
often :: `=v v`	ϵ :: `=t +wh c`

Training data

90	pierre will praise marie
5	pierre will **often** praise marie
1	who pierre will praise
1	who pierre will **often** praise

Maximise likelihood via stochastic gradient ascent:

$$P_{\boldsymbol{\lambda}}(N \to \delta) = \frac{\exp(\boldsymbol{\lambda} \cdot \phi(N \to \delta))}{\sum \exp(\boldsymbol{\lambda} \cdot \phi(N \to \delta'))}$$

	naive	smarter
$st :: \langle \mathtt{v} \rangle_0 \to s :: \langle \mathtt{=d\ v} \rangle_1 \quad t :: \langle \mathtt{d} \rangle_1$	0.95	0.94
$st :: \langle \mathtt{v} \rangle_0 \to s :: \langle \mathtt{=v\ v} \rangle_1 \quad t :: \langle \mathtt{v} \rangle_0$	0.05	0.06
$\langle s, t \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 \to s :: \langle \mathtt{=d\ v} \rangle_1 \quad t :: \langle \mathtt{d\ -wh} \rangle_1$	**0.67**	**0.94**
$\langle st, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0 \to s :: \langle \mathtt{=v\ v} \rangle_1 \quad \langle t, u \rangle :: \langle \mathtt{v}, \mathtt{-wh} \rangle_0$	**0.33**	**0.06**

Naive parametrization

$$
\begin{array}{ccccc}
G_A & \longrightarrow & \text{Naive parametrization} & \longrightarrow & \begin{array}{l}0.95\\0.05\\0.67\\0.33\end{array}\\
 & & \uparrow & & \\
 & & \text{Training corpus} & &
\end{array}
$$

$$
\begin{array}{ccccc}
G_{B1} & \longrightarrow & \text{Naive parametrization} & \longrightarrow & \begin{array}{l}0.48\\0.24\\0.14\\0.10\\0.05\end{array}\\
 & & \uparrow & & \\
 & & \text{Training corpus} & & \\
 & & \downarrow & & \\
G_{B2} & \longrightarrow & \text{Naive parametrization} & \longrightarrow & \begin{array}{l}0.48\\0.24\\0.14\\0.10\\0.05\end{array}
\end{array}
$$

Smarter parametrization

$$
\begin{array}{ccccc}
G_A & \longrightarrow & \text{Smarter parametrization} & \longrightarrow & \begin{array}{l}0.94\\0.06\\0.94\\0.06\end{array}\\
 & & \uparrow & & \\
 & & \text{Training corpus} & &
\end{array}
$$

Solution #2 with the smarter parametrization

$$\begin{tikzcd} \boxed{\text{“normal” MG}} \arrow[r] \arrow[d] & \boxed{\text{MCFG}} \\ \boxed{\text{“re-merge” MG}} \arrow[r] & \boxed{\text{MCFG}} \end{tikzcd}$$

Language of both grammars

boys will shave
boys will shave themselves
who will shave
who will shave themselves
foo boys will shave
foo boys will shave themselves

Training data

10	boys will shave
2	boys will shave themselves
3	who will shave
1	who will shave themselves
5	foo boys will shave

Solution #2 with the smarter parametrization

$$\begin{tikzcd} \boxed{\text{“normal” MG}} \arrow[r] \arrow[d] & \boxed{\text{MCFG}} \\ \boxed{\text{“re-merge” MG}} \arrow[r] & \boxed{\text{MCFG}} \end{tikzcd}$$

Language of both grammars

boys will shave
boys will shave themselves
who will shave
who will shave themselves
foo boys will shave
foo boys will shave themselves

Training data

10	boys will shave
2	boys will shave themselves
3	who will shave
1	who will shave themselves
5	foo boys will shave

Merge and move distinct operations:

0.35478	boys will shave
0.35478	foo boys will shave
0.14801	who will shave
0.05022	boys will shave themselves
0.05022	foo boys will shave themselves
0.04199	who will shave themselves

Solution #2 with the smarter parametrization

$$\begin{tikzcd} \text{``normal'' MG} \arrow[r] \arrow[d] & \text{MCFG} \\ \text{``re-merge'' MG} \arrow[r] & \text{MCFG} \end{tikzcd}$$

Language of both grammars

boys will shave
boys will shave themselves
who will shave
who will shave themselves
foo boys will shave
foo boys will shave themselves

Training data

10	boys will shave
2	boys will shave themselves
3	who will shave
1	who will shave themselves
5	foo boys will shave

Merge and move distinct operations:

0.35478	boys will shave
0.35478	foo boys will shave
0.14801	who will shave
0.05022	boys will shave themselves
0.05022	foo boys will shave themselves
0.04199	who will shave themselves

Merge and move unified:

0.35721	boys will shave
0.35721	foo boys will shave
0.095	who will shave
0.095	who will shave themselves
0.04779	boys will shave themselves
0.04779	foo boys will shave themselves

Solution #2 with the smarter parametrization

$$
\begin{tikzcd}
\text{“normal” MG} \arrow[r] \arrow[d] & \text{MCFG} \\
\text{“re-merge” MG} \arrow[r] & \text{MCFG}
\end{tikzcd}
$$

Language of both grammars

boys will shave
boys will shave themselves
who will shave
who will shave themselves
foo boys will shave
foo boys will shave themselves

Training data

10	boys will shave
2	boys will shave themselves
3	who will shave
1	who will shave themselves
5	foo boys will shave

Merge and move distinct operations:

0.35478	boys will shave
0.35478	foo boys will shave
0.14801	who will shave
0.05022	boys will shave themselves
0.05022	foo boys will shave themselves
0.04199	who will shave themselves

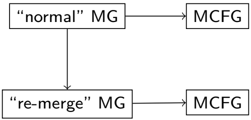
	Entropy	Entropy Reduction
—	2.09	—
who	0.76	1.33
will	0.76	0.00
shave	0.76	0.00
themselves	0.00	0.76

Merge and move unified:

0.35721	boys will shave
0.35721	foo boys will shave
0.095	who will shave
0.095	who will shave themselves
0.04779	boys will shave themselves
0.04779	foo boys will shave themselves

	Entropy	Entropy Reduction
—	2.13	—
who	1.00	1.13
will	1.00	0.00
shave	1.00	0.00
themselves	0.00	1.00

Solution #2 with the smarter parametrization



Language of both grammars

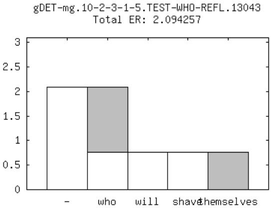
boys will shave
boys will shave themselves
who will shave
who will shave themselves
foo boys will shave
foo boys will shave themselves

Training data

10	boys will shave
2	boys will shave themselves
3	who will shave
1	who will shave themselves
5	foo boys will shave

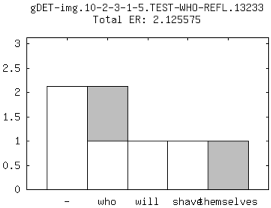
Merge and move distinct operations:

0.35478	boys will shave
0.35478	foo boys will shave
0.14801	who will shave
0.05022	boys will shave themselves
0.05022	foo boys will shave themselves
0.04199	who will shave themselves



Merge and move unified:

0.35721	boys will shave
0.35721	foo boys will shave
0.095	who will shave
0.095	who will shave themselves
0.04779	boys will shave themselves
0.04779	foo boys will shave themselves



Naive parametrization

G_A ⟶ Naive parametrization ⟶ 0.95, 0.05, 0.67, 0.33

Training corpus ⟶ Naive parametrization

G_{B1} ⟶ Naive parametrization ⟶ 0.48, 0.24, 0.14, 0.10, 0.05

G_{B2} ⟶ Naive parametrization ⟶ 0.48, 0.24, 0.14, 0.10, 0.05

Training corpus ⟶ Naive parametrization (G_{B1}), Naive parametrization (G_{B2})

Smarter parametrization

G_A ⟶ Smarter parametrization ⟶ 0.94, 0.06, 0.94, 0.06

Training corpus ⟶ Smarter parametrization

G_{B1} ⟶ Smarter parametrization ⟶ 0.35, 0.35, 0.15, 0.05, 0.05, 0.04

G_{B2} ⟶ Smarter parametrization ⟶ 0.36, 0.36, 0.10, 0.10, 0.05, 0.05

Training corpus ⟶ Smarter parametrization (G_{B1}), Smarter parametrization (G_{B2})

Choice points in the MG-derived MCFG

Question or not?			
$\langle \texttt{c} \rangle_0$	$\rightarrow$	$\langle \texttt{=t c} \rangle_0$	$\langle \texttt{t} \rangle_0$
$\langle \texttt{c} \rangle_0$	$\rightarrow$	$\langle \texttt{+wh c}, \texttt{-wh} \rangle_0$	

Antecedent lexical or complex?			
$\langle \texttt{ant -subj} \rangle_0$	$\rightarrow$	$\langle \texttt{=subj ant -subj} \rangle_1$	$\langle \texttt{subj} \rangle_0$
$\langle \texttt{ant -subj} \rangle_0$	$\rightarrow$	$\langle \texttt{=subj ant -subj} \rangle_1$	$\langle \texttt{subj} \rangle_1$

Non-wh subject merged and complex, merged and lexical, or moved?			
$\langle \texttt{t} \rangle_0$	$\rightarrow$	$\langle \texttt{=subj t} \rangle_0$	$\langle \texttt{subj} \rangle_0$
$\langle \texttt{t} \rangle_0$	$\rightarrow$	$\langle \texttt{=subj t} \rangle_0$	$\langle \texttt{subj} \rangle_1$
$\langle \texttt{t} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj t}, \texttt{-subj} \rangle_0$	

Wh-phrase same as moving subject or separated because of doubling?			
$\langle \texttt{t}, \texttt{-wh} \rangle_0$	$\rightarrow$	$\langle \texttt{=subj t} \rangle_0$	$\langle \texttt{subj -wh} \rangle_1$
$\langle \texttt{t}, \texttt{-wh} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj t}, \texttt{-subj}, \texttt{-wh} \rangle_0$	

Choice points in the MG-derived MCFG

Question or not?				
$\langle\texttt{c}\rangle_0$	$\rightarrow$	$\langle\texttt{=t c}\rangle_0$	$\langle\texttt{t}\rangle_0$	$\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{t}})$
$\langle\texttt{c}\rangle_0$	$\rightarrow$	$\langle\texttt{+wh c}, \texttt{-wh}\rangle_0$		$\exp(\lambda_{\text{MOVE}} + \lambda_{\texttt{wh}})$

Antecedent lexical or complex?				
$\langle\texttt{ant -subj}\rangle_0$	$\rightarrow$	$\langle\texttt{=subj ant -subj}\rangle_1$	$\langle\texttt{subj}\rangle_0$	$\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{subj}})$
$\langle\texttt{ant -subj}\rangle_0$	$\rightarrow$	$\langle\texttt{=subj ant -subj}\rangle_1$	$\langle\texttt{subj}\rangle_1$	$\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{subj}})$

Non-wh subject merged and complex, merged and lexical, or moved?				
$\langle\texttt{t}\rangle_0$	$\rightarrow$	$\langle\texttt{=subj t}\rangle_0$	$\langle\texttt{subj}\rangle_0$	$\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{subj}})$
$\langle\texttt{t}\rangle_0$	$\rightarrow$	$\langle\texttt{=subj t}\rangle_0$	$\langle\texttt{subj}\rangle_1$	$\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{subj}})$
$\langle\texttt{t}\rangle_0$	$\rightarrow$	$\langle\texttt{+subj t}, \texttt{-subj}\rangle_0$		$\exp(\lambda_{\text{MOVE}} + \lambda_{\texttt{subj}})$

Wh-phrase same as moving subject or separated because of doubling?				
$\langle\texttt{t}, \texttt{-wh}\rangle_0$	$\rightarrow$	$\langle\texttt{=subj t}\rangle_0$	$\langle\texttt{subj -wh}\rangle_1$	$\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{subj}})$
$\langle\texttt{t}, \texttt{-wh}\rangle_0$	$\rightarrow$	$\langle\texttt{+subj t}, \texttt{-subj}, \texttt{-wh}\rangle_0$		$\exp(\lambda_{\text{MOVE}} + \lambda_{\texttt{subj}})$

Learned weights on the MG

$$
\begin{aligned}
\lambda_{\texttt{t}} &= 0.094350 & \exp(\lambda_{\texttt{t}}) &= 1.0989 \\
\lambda_{\texttt{subj}} &= -5.734063 & \exp(\lambda_{\texttt{v}}) &= 0.0032 \\
\lambda_{\texttt{wh}} &= -0.094350 & \exp(\lambda_{\texttt{wh}}) &= 0.9100 \\
\lambda_{\text{MERGE}} &= 0.629109 & \exp(\lambda_{\text{MERGE}}) &= 1.8759 \\
\lambda_{\text{MOVE}} &= -0.629109 & \exp(\lambda_{\text{MOVE}}) &= 0.5331
\end{aligned}
$$

Learned weights on the MG

$$\lambda_{\mathtt{t}} = 0.094350 \qquad \exp(\lambda_{\mathtt{t}}) = 1.0989$$
$$\lambda_{\mathtt{subj}} = -5.734063 \qquad \exp(\lambda_{\mathtt{v}}) = 0.0032$$
$$\lambda_{\mathtt{wh}} = -0.094350 \qquad \exp(\lambda_{\mathtt{wh}}) = 0.9100$$
$$\lambda_{\textsc{merge}} = 0.629109 \qquad \exp(\lambda_{\textsc{merge}}) = 1.8759$$
$$\lambda_{\textsc{move}} = -0.629109 \qquad \exp(\lambda_{\textsc{move}}) = 0.5331$$

$$P(\text{antecedent is lexical}) = 0.5$$
$$P(\text{antecedent is non-lexical}) = 0.5$$

$$P(\text{wh-phrase reflexivized}) = \frac{\exp(\lambda_{\textsc{move}})}{\exp(\lambda_{\textsc{merge}}) + \exp(\lambda_{\textsc{move}})} = 0.2213$$
$$P(\text{wh-phrase non-reflexivized}) = \frac{\exp(\lambda_{\textsc{merge}})}{\exp(\lambda_{\textsc{merge}}) + \exp(\lambda_{\textsc{move}})} = 0.7787$$

$$P(\text{question}) = \frac{\exp(\lambda_{\textsc{move}} + \lambda_{\mathtt{wh}})}{\exp(\lambda_{\textsc{merge}} + \lambda_{\mathtt{t}}) + \exp(\lambda_{\textsc{move}} + \lambda_{\mathtt{wh}})} = 0.1905$$
$$P(\text{non-question}) = \frac{\exp(\lambda_{\textsc{merge}} + \lambda_{\mathtt{t}})}{\exp(\lambda_{\textsc{merge}} + \lambda_{\mathtt{t}}) + \exp(\lambda_{\textsc{move}} + \lambda_{\mathtt{wh}})} = 0.8095$$

$$P(\text{non-wh subject merged and complex}) = \frac{\exp(\lambda_{\textsc{merge}})}{\exp(\lambda_{\textsc{merge}}) + \exp(\lambda_{\textsc{merge}}) + \exp(\lambda_{\textsc{move}})} = 0.4378$$
$$P(\text{non-wh subject merged and lexical}) = \frac{\exp(\lambda_{\textsc{merge}})}{\exp(\lambda_{\textsc{merge}}) + \exp(\lambda_{\textsc{merge}}) + \exp(\lambda_{\textsc{move}})} = 0.4378$$
$$P(\text{non-wh subject moved}) = \frac{\exp(\lambda_{\textsc{move}})}{\exp(\lambda_{\textsc{merge}}) + \exp(\lambda_{\textsc{merge}}) + \exp(\lambda_{\textsc{move}})} = 0.1244$$

Learned weights on the MG

$$\lambda_{\texttt{t}} = 0.094350 \qquad \exp(\lambda_{\texttt{t}}) = 1.0989$$
$$\lambda_{\texttt{subj}} = -5.734063 \qquad \exp(\lambda_{\texttt{v}}) = 0.0032$$
$$\lambda_{\texttt{wh}} = -0.094350 \qquad \exp(\lambda_{\texttt{wh}}) = 0.9100$$
$$\lambda_{\text{MERGE}} = 0.629109 \qquad \exp(\lambda_{\text{MERGE}}) = 1.8759$$
$$\lambda_{\text{MOVE}} = -0.629109 \qquad \exp(\lambda_{\text{MOVE}}) = 0.5331$$

$$P(\text{antecedent is lexical}) = 0.5$$
$$P(\text{antecedent is non-lexical}) = 0.5$$

$$P(\text{wh-phrase reflexivized}) = \frac{\exp(\lambda_{\text{MOVE}})}{\exp(\lambda_{\text{MERGE}}) + \exp(\lambda_{\text{MOVE}})} = 0.2213$$
$$P(\text{wh-phrase non-reflexivized}) = \frac{\exp(\lambda_{\text{MERGE}})}{\exp(\lambda_{\text{MERGE}}) + \exp(\lambda_{\text{MOVE}})} = 0.7787$$

$$P(\text{question}) = \frac{\exp(\lambda_{\text{MOVE}} + \lambda_{\texttt{wh}})}{\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{t}}) + \exp(\lambda_{\text{MOVE}} + \lambda_{\texttt{wh}})} = 0.1905$$
$$P(\text{non-question}) = \frac{\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{t}})}{\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{t}}) + \exp(\lambda_{\text{MOVE}} + \lambda_{\texttt{wh}})} = 0.8095$$

$$P(\text{non-wh subject merged and complex}) = \frac{\exp(\lambda_{\text{MERGE}})}{\exp(\lambda_{\text{MERGE}}) + \exp(\lambda_{\text{MERGE}}) + \exp(\lambda_{\text{MOVE}})} = 0.4378$$
$$P(\text{non-wh subject merged and lexical}) = \frac{\exp(\lambda_{\text{MERGE}})}{\exp(\lambda_{\text{MERGE}}) + \exp(\lambda_{\text{MERGE}}) + \exp(\lambda_{\text{MOVE}})} = 0.4378$$
$$P(\text{non-wh subject moved}) = \frac{\exp(\lambda_{\text{MOVE}})}{\exp(\lambda_{\text{MERGE}}) + \exp(\lambda_{\text{MERGE}}) + \exp(\lambda_{\text{MOVE}})} = 0.1244$$

$$P(\text{who will shave}) = 0.1905 \times 0.7787 = 0.148$$
$$P(\text{boys will shave themselves}) = 0.5 \times 0.8095 \times 0.1244 = 0.050$$

Choice points in the IMG-derived MCFG

Question or not?		
$\langle \texttt{-c} \rangle_0$	$\rightarrow$	$\langle \texttt{+t -c}, \texttt{-t} \rangle_1$
$\langle \texttt{-c} \rangle_0$	$\rightarrow$	$\langle \texttt{+wh -c}, \texttt{-wh} \rangle_0$

Antecedent lexical or complex?			
$\langle \texttt{+subj -ant -subj}, \texttt{-subj} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj -ant -subj} \rangle_0$	$\langle \texttt{-subj} \rangle_0$
$\langle \texttt{+subj -ant -subj}, \texttt{-subj} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj -ant -subj} \rangle_0$	$\langle \texttt{-subj} \rangle_1$

Non-wh subject merged and complex, merged and lexical, or moved?			
$\langle \texttt{+subj -t}, \texttt{-subj} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj -t} \rangle_0$	$\langle \texttt{-subj} \rangle_0$
$\langle \texttt{+subj -t}, \texttt{-subj} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj -t} \rangle_0$	$\langle \texttt{-subj} \rangle_1$
$\langle \texttt{+subj -t}, \texttt{-subj} \rangle_0$	$\rightarrow$	$\langle \texttt{+v +subj -t}, \texttt{-v}, \texttt{-subj} \rangle_1$	

Wh-phrase same as moving subject or separated because of doubling?		
$\langle \texttt{-t}, \texttt{-wh} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj -t}, \texttt{-subj -wh} \rangle_0$
$\langle \texttt{-t}, \texttt{-wh} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj -t}, \texttt{-subj}, \texttt{-wh} \rangle_0$

Choice points in the IMG-derived MCFG

Question or not?			
$\langle \texttt{-c} \rangle_0$	$\rightarrow$	$\langle \texttt{+t -c}, \texttt{-t} \rangle_1$	$\exp(\lambda_{\text{MRG}} + \lambda_{\texttt{t}})$
$\langle \texttt{-c} \rangle_0$	$\rightarrow$	$\langle \texttt{+wh -c}, \texttt{-wh} \rangle_0$	$\exp(\lambda_{\text{MRG}} + \lambda_{\texttt{wh}})$

Antecedent lexical or complex?				
$\langle \texttt{+subj -ant -subj}, \texttt{-subj} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj -ant -subj} \rangle_0$	$\langle \texttt{-subj} \rangle_0$	$\exp(\lambda_{\text{INSERT}})$
$\langle \texttt{+subj -ant -subj}, \texttt{-subj} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj -ant -subj} \rangle_0$	$\langle \texttt{-subj} \rangle_1$	$\exp(\lambda_{\text{INSERT}})$

Non-wh subject merged and complex, merged and lexical, or moved?				
$\langle \texttt{+subj -t}, \texttt{-subj} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj -t} \rangle_0$	$\langle \texttt{-subj} \rangle_0$	$\exp(\lambda_{\text{INSERT}})$
$\langle \texttt{+subj -t}, \texttt{-subj} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj -t} \rangle_0$	$\langle \texttt{-subj} \rangle_1$	$\exp(\lambda_{\text{INSERT}})$
$\langle \texttt{+subj -t}, \texttt{-subj} \rangle_0$	$\rightarrow$	$\langle \texttt{+v +subj -t}, \texttt{-v}, \texttt{-subj} \rangle_1$		$\exp(\lambda_{\text{MRG}} + \lambda_{\texttt{v}})$

Wh-phrase same as moving subject or separated because of doubling?			
$\langle \texttt{-t}, \texttt{-wh} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj -t}, \texttt{-subj -wh} \rangle_0$	$\exp(\lambda_{\text{MRG}} + \lambda_{\texttt{subj}})$
$\langle \texttt{-t}, \texttt{-wh} \rangle_0$	$\rightarrow$	$\langle \texttt{+subj -t}, \texttt{-subj}, \texttt{-wh} \rangle_0$	$\exp(\lambda_{\text{MRG}} + \lambda_{\texttt{subj}})$

Learned weights on the IMG

$$\begin{aligned}
\lambda_{\mathtt{t}} &= 0.723549 & \exp(\lambda_{\mathtt{t}}) &= 2.0617 \\
\lambda_{\mathtt{v}} &= 0.440585 & \exp(\lambda_{\mathtt{v}}) &= 1.5536 \\
\lambda_{\mathtt{wh}} &= -0.723459 & \exp(\lambda_{\mathtt{wh}}) &= 0.4850 \\
\lambda_{\mathrm{INSERT}} &= 0.440585 & \exp(\lambda_{\mathrm{INSERT}}) &= 1.5536 \\
\lambda_{\mathrm{MRG}} &= -0.440585 & \exp(\lambda_{\mathrm{MRG}}) &= 0.6437
\end{aligned}$$

Learned weights on the IMG

$$\lambda_{\mathtt{t}} = 0.723549 \qquad \exp(\lambda_{\mathtt{t}}) = 2.0617$$
$$\lambda_{\mathtt{v}} = 0.440585 \qquad \exp(\lambda_{\mathtt{v}}) = 1.5536$$
$$\lambda_{\mathtt{wh}} = -0.723459 \qquad \exp(\lambda_{\mathtt{wh}}) = 0.4850$$
$$\lambda_{\text{INSERT}} = 0.440585 \qquad \exp(\lambda_{\text{INSERT}}) = 1.5536$$
$$\lambda_{\text{MRG}} = -0.440585 \qquad \exp(\lambda_{\text{MRG}}) = 0.6437$$

$$P(\text{antecedent is lexical}) = 0.5$$
$$P(\text{antecedent is non-lexical}) = 0.5$$

$$P(\text{wh-phrase reflexivized}) = 0.5$$
$$P(\text{wh-phrase non-reflexivized}) = 0.5$$

$$P(\text{question}) = \frac{\exp(\lambda_{\text{MRG}} + \lambda_{\mathtt{wh}})}{\exp(\lambda_{\text{MRG}} + \lambda_{\mathtt{t}}) + \exp(\lambda_{\text{MRG}} + \lambda_{\mathtt{wh}})} = \frac{\exp(\lambda_{\mathtt{wh}})}{\exp(\lambda_{\mathtt{t}}) + \exp(\lambda_{\mathtt{wh}})} = 0.1905$$
$$P(\text{non-question}) = \frac{\exp(\lambda_{\text{MRG}} + \lambda_{\mathtt{t}})}{\exp(\lambda_{\text{MRG}} + \lambda_{\mathtt{t}}) + \exp(\lambda_{\text{MRG}} + \lambda_{\mathtt{wh}})} = \frac{\exp(\lambda_{\mathtt{t}})}{\exp(\lambda_{\mathtt{t}}) + \exp(\lambda_{\mathtt{wh}})} = 0.8095$$

$$P(\text{non-wh subject merged and lexical}) = \frac{\exp(\lambda_{\text{INSERT}})}{\exp(\lambda_{\text{INSERT}}) + \exp(\lambda_{\text{INSERT}}) + \exp(\lambda_{\text{MRG}} + \lambda_{\mathtt{v}})} = 0.4412$$
$$P(\text{non-wh subject merged and complex}) = \frac{\exp(\lambda_{\text{INSERT}})}{\exp(\lambda_{\text{INSERT}}) + \exp(\lambda_{\text{INSERT}}) + \exp(\lambda_{\text{MRG}} + \lambda_{\mathtt{v}})} = 0.4412$$
$$P(\text{non-wh subject moved}) = \frac{\exp(\lambda_{\text{MRG}} + \lambda_{\mathtt{v}})}{\exp(\lambda_{\text{INSERT}}) + \exp(\lambda_{\text{INSERT}}) + \exp(\lambda_{\text{MRG}} + \lambda_{\mathtt{v}})} = 0.1176$$

Learned weights on the IMG

$$\begin{aligned} \lambda_{\mathtt{t}} &= 0.723549 & \exp(\lambda_{\mathtt{t}}) &= 2.0617 \\ \lambda_{\mathtt{v}} &= 0.440585 & \exp(\lambda_{\mathtt{v}}) &= 1.5536 \\ \lambda_{\mathtt{wh}} &= -0.723459 & \exp(\lambda_{\mathtt{wh}}) &= 0.4850 \\ \lambda_{\mathrm{INSERT}} &= 0.440585 & \exp(\lambda_{\mathrm{INSERT}}) &= 1.5536 \\ \lambda_{\mathrm{MRG}} &= -0.440585 & \exp(\lambda_{\mathrm{MRG}}) &= 0.6437 \end{aligned}$$

$$\begin{aligned} P(\text{antecedent is lexical}) &= 0.5 \\ P(\text{antecedent is non-lexical}) &= 0.5 \end{aligned}$$

$$\begin{aligned} P(\text{wh-phrase reflexivized}) &= 0.5 \\ P(\text{wh-phrase non-reflexivized}) &= 0.5 \end{aligned}$$

$$P(\text{question}) = \frac{\exp(\lambda_{\mathrm{MRG}} + \lambda_{\mathtt{wh}})}{\exp(\lambda_{\mathrm{MRG}} + \lambda_{\mathtt{t}}) + \exp(\lambda_{\mathrm{MRG}} + \lambda_{\mathtt{wh}})} = \frac{\exp(\lambda_{\mathtt{wh}})}{\exp(\lambda_{\mathtt{t}}) + \exp(\lambda_{\mathtt{wh}})} = 0.1905$$

$$P(\text{non-question}) = \frac{\exp(\lambda_{\mathrm{MRG}} + \lambda_{\mathtt{t}})}{\exp(\lambda_{\mathrm{MRG}} + \lambda_{\mathtt{t}}) + \exp(\lambda_{\mathrm{MRG}} + \lambda_{\mathtt{wh}})} = \frac{\exp(\lambda_{\mathtt{t}})}{\exp(\lambda_{\mathtt{t}}) + \exp(\lambda_{\mathtt{wh}})} = 0.8095$$

$$P(\text{non-wh subject merged and lexical}) = \frac{\exp(\lambda_{\mathrm{INSERT}})}{\exp(\lambda_{\mathrm{INSERT}}) + \exp(\lambda_{\mathrm{INSERT}}) + \exp(\lambda_{\mathrm{MRG}} + \lambda_{\mathtt{v}})} = 0.4412$$

$$P(\text{non-wh subject merged and complex}) = \frac{\exp(\lambda_{\mathrm{INSERT}})}{\exp(\lambda_{\mathrm{INSERT}}) + \exp(\lambda_{\mathrm{INSERT}}) + \exp(\lambda_{\mathrm{MRG}} + \lambda_{\mathtt{v}})} = 0.4412$$

$$P(\text{non-wh subject moved}) = \frac{\exp(\lambda_{\mathrm{MRG}} + \lambda_{\mathtt{v}})}{\exp(\lambda_{\mathrm{INSERT}}) + \exp(\lambda_{\mathrm{INSERT}}) + \exp(\lambda_{\mathrm{MRG}} + \lambda_{\mathtt{v}})} = 0.1176$$

$$P(\text{who will shave}) = 0.5 \times 0.1905 = 0.095$$

$$P(\text{boys will shave themselves}) = 0.5 \times 0.8095 \times 0.1176 = 0.048$$

Billot, S. and Lang, B. (1989). The structure of shared forests in ambiguous parsing. In *Proceedings of the 1989 Meeting of the Association of Computational Linguistics*.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Chomsky, N. (1980). *Rules and Representations*. Columbia University Press, New York.

Ferreira, F. (2005). Psycholinguistics, formal grammars, and cognitive science. *The Linguistic Review*, 22:365–380.

Gärtner, H.-M. and Michaelis, J. (2010). On the Treatment of Multiple-Wh Interrogatives in Minimalist Grammars. In Hanneforth, T. and Fanselow, G., editors, *Language and Logos*, pages 339–366. Akademie Verlag, Berlin.

Gibson, E. and Wexler, K. (1994). Triggers. *Linguistic Inquiry*, 25:407–454.

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, 30:643–Âŋ672.

Hale, J. T. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Hunter, T. (2011). Insertion Minimalist Grammars: Eliminating redundancies between merge and move. In Kanazawa, M., Kornai, A., Kracht, M., and Seki, H., editors, *The Mathematics of Language (MOL 12 Proceedings)*, volume 6878 of *LNCS*, pages 90–107, Berlin Heidelberg. Springer.

Hunter, T. and Dyer, C. (2013). Distributions on minimalist grammar derivations. In *Proceedings of the 13th Meeting on the Mathematics of Language*.

Koopman, H. and Szabolcsi, A. (2000). *Verbal Complexes*. MIT Press, Cambridge, MA.

Lang, B. (1988). Parsing incomplete sentences. In *Proceedings of the 12th International Conference on Computational Linguistics*, pages 365–371.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Michaelis, J. (2001). Derivational minimalism is mildly context-sensitive. In Moortgat, M., editor, *Logical Aspects of Computational Linguistics*, volume 2014 of *LNCS*, pages 179–198. Springer, Berlin Heidelberg.

Miller, G. A. and Chomsky, N. (1963). Finitary models of language users. In Luce, R. D., Bush, R. R., and Galanter, E., editors, *Handbook of Mathematical Psychology*, volume 2. Wiley and Sons, New York.

Morrill, G. (1994). *Type Logical Grammar: Categorial Logic of Signs*. Kluwer, Dordrecht.

Nederhof, M. J. and Satta, G. (2008). Computing partition functions of pcfgs. *Research on Language and Computation*, 6(2):139–162.

Seki, H., Matsumara, T., Fujii, M., and Kasami, T. (1991). On multiple context-free grammars. *Theoretical Computer Science*, 88:191–229.

Stabler, E. P. (2006). Sidewards without copying. In Wintner, S., editor, *Proceedings of The 11th Conference on Formal Grammar*, pages 157–170, Stanford, CA. CSLI Publications.

Stabler, E. P. (2011). Computational perspectives on minimalism. In Boeckx, C., editor, *The Oxford Handbook of Linguistic Minimalism*. Oxford University Press, Oxford.

Stabler, E. P. and Keenan, E. L. (2003). Structural similarity within and among languages. *Theoretical Computer Science*, 293:345–363.

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

Weir, D. (1988). *Characterizing mildly context-sensitive grammar formalisms*. PhD thesis, University of Pennsylvania.

Yngve, V. H. (1960). A model and an hypothesis for language structure. In *Proceedings of the American Philosophical Society*, volume 104, pages 444–466.