The empirical significance of derivational operations

Tim Hunter

University of California, Los Angeles

February 15, 2017

Overview

Take-home message

Distinct hypotheses about derivational operations can have distinct consequences for our theories' predictions about speakers' linguistic behavior.

Overview

Take-home message

Distinct hypotheses about derivational operations can have distinct consequences for our theories' predictions about speakers' linguistic behavior.

Illustrative case study:

- distinct implementations of movement: move vs. remerge
- distinct predictions when plugged into models of
  - sentence comprehension difficulty, via surprisal
  - grammar selection by a learner, via simple maximum likelihood learning

Overview

Take-home message

Distinct hypotheses about derivational operations can have distinct consequences for our theories' predictions about speakers' linguistic behavior.

Illustrative case study:

- distinct implementations of movement: move vs. remerge
- distinct predictions when plugged into models of
  - sentence comprehension difficulty, via surprisal
  - grammar selection by a learner, via simple maximum likelihood learning

Key idea: Derivations encode the relationship between

- primitive, memorized chunks of knowledge; and
- consequences computed from those

Outline

Outline

Grammars

Q: What is a grammar?
A: A (finite) collection of memorized statements that allows a speaker to recognize an (infinite) collection of expressions

Grammars

Q: What is a grammar?
A: A (finite) collection of memorized statements that allows a speaker to recognize an (infinite) collection of expressions

Why do we want a grammar rather than just a list of expressions?

- The infinitely many expressions can't be encoded directly in a finite mind
- We recognize as well-formed (and understand, and . . .) expressions we've never encountered before

Grammars

Q: What is a grammar?
A: A (finite) collection of memorized statements that allows a speaker to recognize an (infinite) collection of expressions

Why do we want a grammar rather than just a list of expressions?

- The infinitely many expressions can't be encoded directly in a finite mind
- We recognize as well-formed (and understand, and ...) expressions we've never encountered before

Q: What is a derivation?
A: A particular chaining-together of interacting, individually-memorized chunks.

Derivations

Q: What is a derivation?

A: A particular chaining-together of interacting, individually-memorized chunks.

Derivations

Q: What is a derivation?
A: A particular chaining-together of interacting, individually-memorized chunks.

How does a speaker recognize ‘painters’ as well-formed?

Derivations

Q: What is a derivation?
A: A particular chaining-together of interacting, individually-memorized chunks.

How does a speaker recognize ‘painters’ as well-formed?

A speaker recognizes ‘painters’ as a well-formed expression by

- knowing that N + ‘-s’ is well-formed if N is well-formed
- recognizing that ‘painter’ is well-formed

Derivations

Q: What is a derivation?
A: A particular chaining-together of interacting, individually-memorized chunks.

How does a speaker recognize ‘painters’ as well-formed?

A speaker recognizes ‘painters’ as a well-formed expression by

- knowing that N + ‘-s’ is well-formed if N is well-formed
- recognizing that ‘painter’ is well-formed

A speaker recognizes ‘painter’ as a well-formed expression by

- knowing that V + ‘-er’ is well-formed if V is well-formed
- knowing that ‘paint’ is well-formed

Derivations

Q: What is a derivation?
A: A particular chaining-together of interacting, individually-memorized chunks.

How does a speaker recognize ‘painters’ as well-formed?

A speaker recognizes ‘painters’ as a well-formed expression by

- knowing that N + ‘-s’ is well-formed if N is well-formed (a rule)
- recognizing that ‘painter’ is well-formed

A speaker recognizes ‘painter’ as a well-formed expression by

- knowing that V + ‘-er’ is well-formed if V is well-formed (a rule)
- knowing that ‘paint’ is well-formed (a lexical item)

Phillips and Lewis (2013)

Three views of derivations: literalist, formalist, extensionalist

These correspond to different views about which linking hypotheses expose the the derivational claims of a theory to empirical testing.

Phillips and Lewis (2013)

Three views of derivations: literalist, formalist, extensionalist

These correspond to different views about which linking hypotheses expose the the derivational claims of a theory to empirical testing.

- Literalist (one extreme):
  Derivational operations are real-time operations, so observing real-time operations bears directly on derivations.

Phillips and Lewis (2013)

Three views of derivations: literalist, formalist, extensionalist

These correspond to different views about which linking hypotheses expose the the derivational claims of a theory to empirical testing.

- Literalist (one extreme):
  Derivational operations are real-time operations, so observing real-time operations bears directly on derivations.
- Extensionalist (other extreme):
  Derivational operations are abstract to a degree that makes no linking hypotheses available.

Phillips and Lewis (2013)

Three views of derivations: literalist, formalist, extensionalist

These correspond to different views about which linking hypotheses expose the the derivational claims of a theory to empirical testing.

- Literalist (one extreme):
  Derivational operations are real-time operations, so observing real-time operations bears directly on derivations.
- Extensionalist (other extreme):
  Derivational operations are abstract to a degree that makes no linking hypotheses available.
- Formalist (middle ground):
  Derivational operations are cognitive hypotheses testable by certain (less direct) linking hypotheses.

Phillips and Lewis (2013)

When we are told, for example, that the wh-word 'what' is initially merged with a verb and subsequently moved to a left peripheral position in the clause, what claim is this making about the human language system?

Phillips and Lewis (2013)

Phillips and Lewis (2013)

> *When we are told, for example, that the wh-word 'what' is initially merged with a verb and subsequently moved to a left peripheral position in the clause, what claim is this making about the human language system?*
>
> *Phillips and Lewis (2013)*

My answer:

- One component of the human language system is the knowledge of a systematic relationship ("merge") that holds between
  - the well-formedness of the expression 'ate what', and
  - the well-formedness of the expressions 'ate' and 'what'.

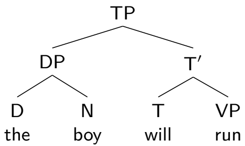
Phillips and Lewis (2013)

> *When we are told, for example, that the wh-word 'what' is initially merged with a verb and subsequently moved to a left peripheral position in the clause, what claim is this making about the human language system?*
>
> *Phillips and Lewis (2013)*

My answer:

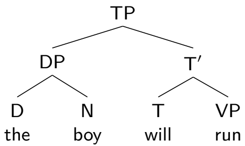
- One component of the human language system is the knowledge of a systematic relationship ("merge") that holds between
  - the well-formedness of the expression 'ate what', and
  - the well-formedness of the expressions 'ate' and 'what'.
- One component of the human language system is the knowledge of a systematic relationship ("move") that holds between
  - the well-formedness of the expression 'what John ate t_i', and
  - the well-formedness of the expression 'John ate what'.

Contemporary syntactic derivations



How does a speaker recognize this as a well-formed expression?

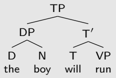
Contemporary syntactic derivations



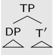
How does a speaker recognize this as a well-formed expression?

Wrong answer: By knowing that this is a well-formed expression.

So some chaining-together of other chunks of knowledge is required, i.e. a derivation.

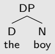
A speaker recognizes that 

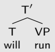
is a well-formed expression by

- knowing that 

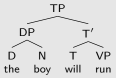
is well-formed if 

and 

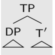
are well-formed
- recognizing that 

is well-formed
- recognizing that 

is well-formed

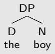
A speaker recognizes that 

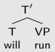
is a well-formed expression by

- knowing that 

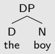
is well-formed if 

and 

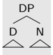
are well-formed
- recognizing that 

is well-formed
- recognizing that 

is well-formed

A speaker recognizes that 

is a well-formed expression by

- knowing that 

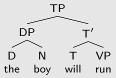
is well-formed if 

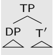
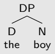
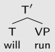
and 

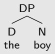
are well-formed
- knowing that 

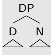
is well-formed
- knowing that 

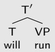
is well-formed

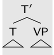
A speaker recognizes that  is a well-formed expression by

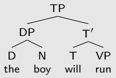
- knowing that  is well-formed if  and  are well-formed
- recognizing that  is well-formed
- recognizing that  is well-formed

A speaker recognizes that  is a well-formed expression by

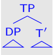
- knowing that  is well-formed if  and  are well-formed
- knowing that  is well-formed
- knowing that  is well-formed

A speaker recognizes that  is a well-formed expression by

- knowing that  is well-formed if  and  are well-formed
- knowing that  is well-formed
- knowing that  is well-formed

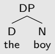
A speaker recognizes that 

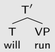
is a well-formed expression by

- knowing that 

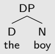
is well-formed if 

and 

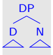
are well-formed (MERGE)
- recognizing that 

is well-formed
- recognizing that 

is well-formed

A speaker recognizes that 

is a well-formed expression by

- knowing that 

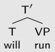
is well-formed if 

and 

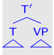
are well-formed (MERGE)
- knowing that 

is well-formed
- knowing that 

is well-formed

A speaker recognizes that 

is a well-formed expression by

- knowing that 

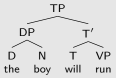
is well-formed if 

and 

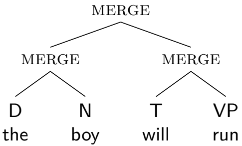
are well-formed (MERGE)
- knowing that 

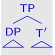
is well-formed
- knowing that 

is well-formed

A speaker recognizes that 

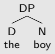
is a well-formed expression by

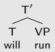


- knowing that 

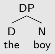
is well-formed if 

and 

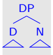
are well-formed (MERGE)
- recognizing that 

is well-formed
- recognizing that 

is well-formed

A speaker recognizes that 

is a well-formed expression by

- knowing that 

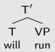
is well-formed if 

and 

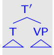
are well-formed (MERGE)
- knowing that 

is well-formed
- knowing that 

is well-formed

A speaker recognizes that 

is a well-formed expression by

- knowing that 

is well-formed if 

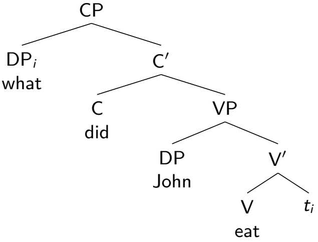
and 

are well-formed (MERGE)
- knowing that 

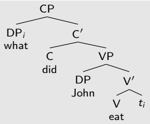
is well-formed
- knowing that 

is well-formed

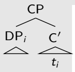
How does a speaker recognize this as a well-formed expression?

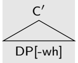


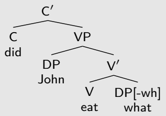
(Same wrong answer as before . . .)

A speaker recognizes that 

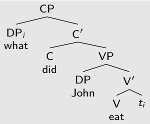
is a well-formed expression by

- knowing that 

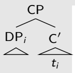
is well-formed if 

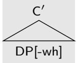
is well-formed
- recognizing that 

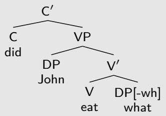
is well-formed

A speaker recognizes that 

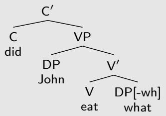
is a well-formed expression by

- knowing that 

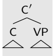
is well-formed if 

is well-formed
- recognizing that 

is well-formed

A speaker recognizes that 

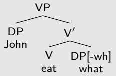
is a well-formed expression by

- knowing that 

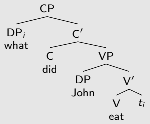
is well-formed if 

and 

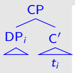
are well-formed
- knowing that 

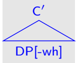
is well-formed
- recognizing that 

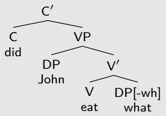
is well-formed

A speaker recognizes that 

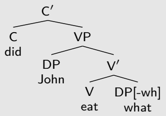
is a well-formed expression by

- knowing that 

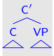
is well-formed if 

is well-formed (MOVE)
- recognizing that 

is well-formed

A speaker recognizes that 

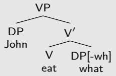
is a well-formed expression by

- knowing that 

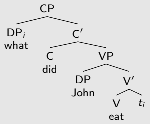
is well-formed if 

and 

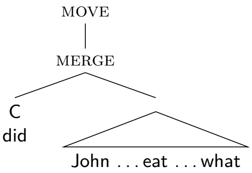
are well-formed (MERGE)
- knowing that 

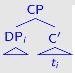
is well-formed
- recognizing that 

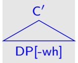
is well-formed

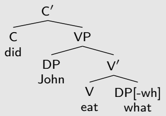
A speaker recognizes that 

is a well-formed expression by

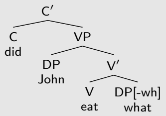


- knowing that 

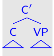
is well-formed if 

is well-formed (MOVE)
- recognizing that 

is well-formed

A speaker recognizes that 

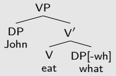
is a well-formed expression by

- knowing that 

is well-formed if 

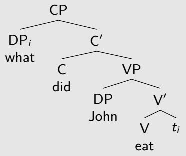
and 

are well-formed (MERGE)
- knowing that 

is well-formed
- recognizing that 

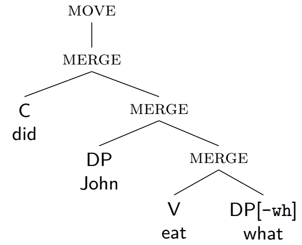
is well-formed

Chaining together memorized chunks

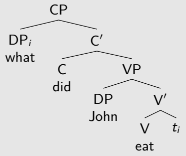
A speaker recognizes that 

is a well-formed expression by . . .

. . . identifying these relationships between memorized chunks of knowledge:

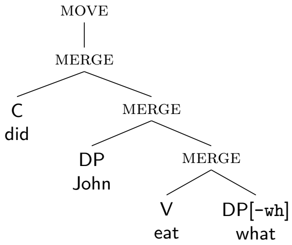


Chaining together memorized chunks

A speaker recognizes that 

is a well-formed expression by . . .

. . . identifying these relationships between memorized chunks of knowledge:



But how can we investigate the chained-together chunks rather than just their results?

Distinct divisions of labour

Suppose we have a black box that recognizes a triple of numbers iff each number is drawn from $\{1, 2, 3, 4, 5, 6\}$.

We have two hypotheses about the "grammar" inside the black box.

Distinct divisions of labour

Suppose we have a black box that recognizes a triple of numbers iff each number is drawn from $\{1, 2, 3, 4, 5, 6\}$.

We have two hypotheses about the "grammar" inside the black box.

Hypothesis #1 involves a blue die and a red die:

- a triple (i, j, k) is well-formed if
  i is well-formed, j is well-formed, and k is well-formed

Distinct divisions of labour

Suppose we have a black box that recognizes a triple of numbers iff each number is drawn from $\{1, 2, 3, 4, 5, 6\}$.

We have two hypotheses about the "grammar" inside the black box.

Hypothesis #1 involves a blue die and a red die:

- a triple (i, j, k) is well-formed if
  i is well-formed, j is well-formed, and k is well-formed

Hypothesis #2 involves a green die and a yellow die:

- a triple (i, j, k) is well-formed if
  i is well-formed, j is well-formed, and k is well-formed

Distinct divisions of labour

Suppose we have a black box that recognizes a triple of numbers iff each number is drawn from $\{1, 2, 3, 4, 5, 6\}$.

We have two hypotheses about the "grammar" inside the black box.

Hypothesis #1 involves a blue die and a red die:

- a triple (i, j, k) is well-formed if
  i is well-formed, j is well-formed, and k is well-formed

Hypothesis #2 involves a green die and a yellow die:

- a triple (i, j, k) is well-formed if
  i is well-formed, j is well-formed, and k is well-formed

These are different "divisions of labour", ways of breaking down the work into (finitely many) chainable chunks

Division of labour

What do these hypotheses say about the triple $(4, 5, 6)$?

Hypothesis #1:

$(4, 5, 6)$ is well-formed if

- 4 is well-formed
- 5 is well-formed
- 6 is well-formed

Hypothesis #2:

$(4, 5, 6)$ is well-formed if

- 4 is well-formed
- 5 is well-formed
- 6 is well-formed

Division of labour

What do these hypotheses say about the triple $(4, 5, 6)$?

Hypothesis #1:

$(4, 5, 6)$ is possible to the extent that

- 4 is possible
- 5 is possible
- 6 is possible

Hypothesis #2:

$(4, 5, 6)$ is possible to the extent that

- 4 is possible
- 5 is possible
- 6 is possible

Division of labour

What do these hypotheses say about the triple $(4, 5, 6)$?

Hypothesis #1:

$(4, 5, 6)$ is probable to the extent that

- 4 is probable
- 5 is probable
- 6 is probable

Hypothesis #2:

$(4, 5, 6)$ is probable to the extent that

- 4 is probable
- 5 is probable
- 6 is probable

Division of labour

What do these hypotheses say about the triple $(4, 5, 6)$?

Hypothesis #1:

$(4, 5, 6)$ is probable to the extent that

- $\boxed{4}$ is probable
- $\boxed{5}$ is probable
- $\boxed{6}$ is probable

So it should pattern with $(5, 4, 6)$.
Both have probability: $P(\boxed{4}) \cdot P(\boxed{5}) \cdot P(\boxed{6})$

Hypothesis #2:

$(4, 5, 6)$ is probable to the extent that

- $\boxed{4}$ is probable
- $\boxed{5}$ is probable
- $\boxed{6}$ is probable

So it should pattern with $(4, 6, 5)$.
Both have probability: $P(\boxed{4}) \cdot P(\boxed{5}) \cdot P(\boxed{6})$

Division of labour

What do these hypotheses say about the triple $(4, 5, 6)$?

Hypothesis #1:

$(4, 5, 6)$ is probable to the extent that

- 4 is probable
- 5 is probable
- 6 is probable

So it should pattern with $(5, 4, 6)$.
Both have probability: $P(\boxed{4}) \cdot P(\boxed{5}) \cdot P(\boxed{6})$

Hypothesis #2:

$(4, 5, 6)$ is probable to the extent that

- 4 is probable
- 5 is probable
- 6 is probable

So it should pattern with $(4, 6, 5)$.
Both have probability: $P(\boxed{4}) \cdot P(\boxed{5}) \cdot P(\boxed{6})$

- Same possibilities: both recognize the set $\{1, 2, 3, 4, 5, 6\}^3$
- Different probabilities: distinct ranges of probability distributions over this set
- More generally: distinct similarity relations over this set

Outline

Roadmap

Plan for this section and the next:

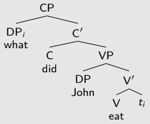
- Introduce two grammatical systems that differ only in their derivational operations (i.e. their "chainable" chunks)
  - merge and move as distinct derivational primitives (Stabler 1997, Keenan and Stabler 2003)
  - merge and move implemented by a single derivational primitive (Stabler 2006, Hunter 2011)

Roadmap

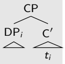
Plan for this section and the next:

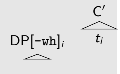
- Introduce two grammatical systems that differ only in their derivational operations (i.e. their "chainable" chunks)
  - merge and move as distinct derivational primitives (Stabler 1997, Keenan and Stabler 2003)
  - merge and move implemented by a single derivational primitive (Stabler 2006, Hunter 2011)
- Show that they produce different empirical predictions when plugged in to common probabilistic modeling settings
  - sentence comprehension difficulty via surprisal
  - selection among candidate grammars by a learner

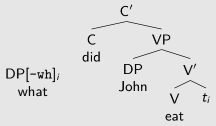
IMG derivations

A speaker recognizes that 

is a well-formed expression by

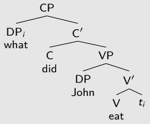
- knowing that 

is well-formed if 

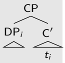
is well-formed (MRG)
- recognizing that 

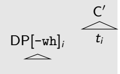
is well-formed

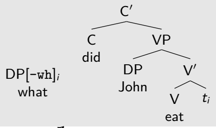
IMG derivations

A speaker recognizes that 

is a well-formed expression by

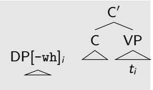
- knowing that 

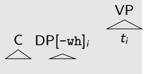
is well-formed if 

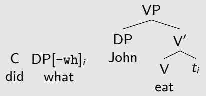
is well-formed (MRG)
- recognizing that 

is well-formed

A speaker recognizes that this is a well-formed expression by

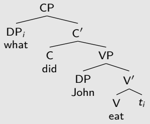
- knowing that 

is well-formed if 

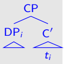
is well-formed (MRG)
- recognizing that 

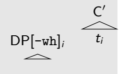
is well-formed

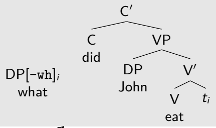
IMG derivations

A speaker recognizes that 

is a well-formed expression by

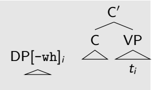
- knowing that 

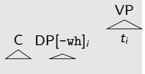
is well-formed if 

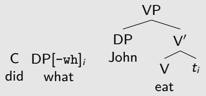
is well-formed (MRG)
- recognizing that 

is well-formed

A speaker recognizes that this is a well-formed expression by

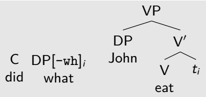
- knowing that 

is well-formed if 

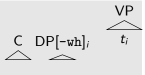
is well-formed (MRG)
- recognizing that 

is well-formed

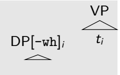
IMG derivations

A speaker recognizes that 

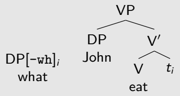
is well-formed by

- knowing that 

is well-formed if 

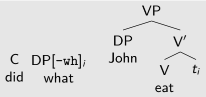
and 

are well-formed (INSERT)
- knowing that 

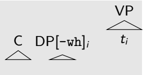
is well-formed
- recognizing that 

is well-formed

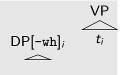
IMG derivations

A speaker recognizes that 

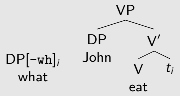
is well-formed by

- knowing that 

is well-formed if 

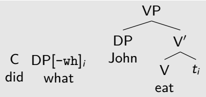
and 

are well-formed (INSERT)
- knowing that 

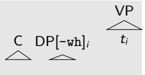
is well-formed
- recognizing that 

is well-formed

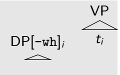
IMG derivations

A speaker recognizes that 

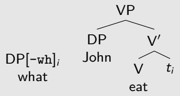
is well-formed by

- knowing that 

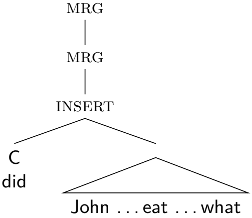
is well-formed if 

and 

are well-formed (INSERT)
- knowing that 

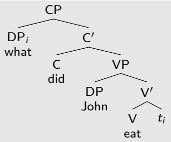
is well-formed
- recognizing that 

is well-formed



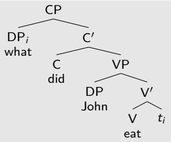
Distinct derivations

Which primitive chunks of knowledge are chained together to produce this expression?

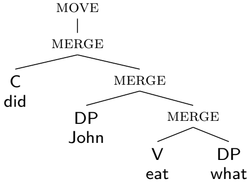


Distinct derivations

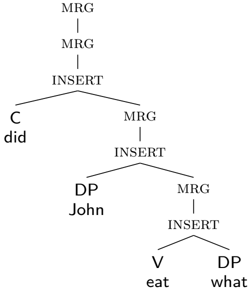
Which primitive chunks of knowledge are chained together to produce this expression?



MG hypothesis: MERGE and MOVE



IMG hypothesis: MRG and INSERT



(NB: Don't be distracted by the difference in the **number** of steps.)

Phillips and Lewis (2013)

When we are told, for example, that the wh-word 'what' is initially merged with a verb and subsequently moved to a left peripheral position in the clause, what claim is this making about the human language system?

Phillips and Lewis (2013)

Phillips and Lewis (2013)

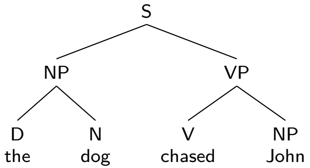
When we are told, for example, that the wh-word 'what' is initially merged with a verb and subsequently moved to a left peripheral position in the clause, what claim is this making about the human language system?

Phillips and Lewis (2013)

When we are told, for example, that the wh-word 'what' is initially inserted and subsequently merged with a verb and then merged into a left peripheral position in the clause, what claim is this making about the human language system?

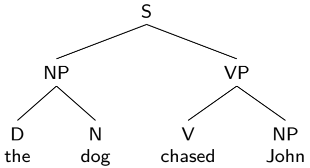
Outline

Probabilities on a CFG



$$
\begin{array}{lllll}
1.0 & \mathsf{S} & \to & \mathsf{NP\ VP} \\
0.3 & \mathsf{NP} & \to & \mathsf{John} \\
0.2 & \mathsf{NP} & \to & \mathsf{he} \\
0.5 & \mathsf{NP} & \to & \mathsf{D\ N} \\
0.3 & \mathsf{D} & \to & \mathsf{the} \\
0.7 & \mathsf{D} & \to & \mathsf{a} \\
0.6 & \mathsf{N} & \to & \mathsf{dog} \\
0.4 & \mathsf{N} & \to & \mathsf{cat} \\
0.8 & \mathsf{VP} & \to & \mathsf{V\ NP} \\
0.2 & \mathsf{VP} & \to & \mathsf{V} \\
0.9 & \mathsf{V} & \to & \mathsf{chased} \\
0.1 & \mathsf{V} & \to & \mathsf{ate}
\end{array}
$$

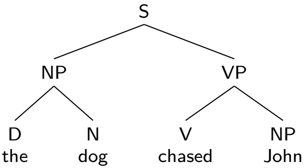
Probabilities on a CFG



1.0	S	$\rightarrow$	NP VP
0.3	NP	$\rightarrow$	John
0.2	NP	$\rightarrow$	he
0.5	NP	$\rightarrow$	D N
0.3	D	$\rightarrow$	the
0.7	D	$\rightarrow$	a
0.6	N	$\rightarrow$	dog
0.4	N	$\rightarrow$	cat
0.8	VP	$\rightarrow$	V NP
0.2	VP	$\rightarrow$	V
0.9	V	$\rightarrow$	chased
0.1	V	$\rightarrow$	ate

$$P(T) = 1.0 \times (0.5 \times 0.3 \times 0.6) \times (0.8 \times 0.9 \times 0.3)$$

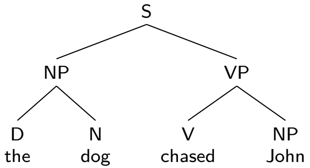
Probabilities on a CFG



$$P(T|\boldsymbol{\lambda}) = \lambda_1 \times (\lambda_4 \times \lambda_5 \times \lambda_7) \times (\lambda_9 \times \lambda_{11} \times \lambda_2)$$

λ_1	S	$\rightarrow$	NP VP
λ_2	NP	$\rightarrow$	John
λ_3	NP	$\rightarrow$	he
λ_4	NP	$\rightarrow$	D N
λ_5	D	$\rightarrow$	the
λ_6	D	$\rightarrow$	a
λ_7	N	$\rightarrow$	dog
λ_8	N	$\rightarrow$	cat
λ_9	VP	$\rightarrow$	V NP
λ_{10}	VP	$\rightarrow$	V
λ_{11}	V	$\rightarrow$	chased
λ_{12}	V	$\rightarrow$	ate

Probabilities on a CFG



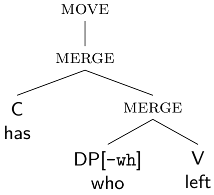
λ_1	S	$\rightarrow$	NP VP
λ_2	NP	$\rightarrow$	John
λ_3	NP	$\rightarrow$	he
λ_4	NP	$\rightarrow$	D N
λ_5	D	$\rightarrow$	the
λ_6	D	$\rightarrow$	a
λ_7	N	$\rightarrow$	dog
λ_8	N	$\rightarrow$	cat
λ_9	VP	$\rightarrow$	V NP
λ_{10}	VP	$\rightarrow$	V
λ_{11}	V	$\rightarrow$	chased
λ_{12}	V	$\rightarrow$	ate

$$P(T|\boldsymbol{\lambda}) = \lambda_1 \times (\lambda_4 \times \lambda_5 \times \lambda_7) \times (\lambda_9 \times \lambda_{11} \times \lambda_2)$$

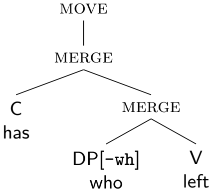
Training question: What values of $\lambda_1, \lambda_2, \ldots, \lambda_{12}$ maximize the likelihood of the training data $P(D|\boldsymbol{\lambda})$?

Note that the choice of grammatical rules (division of labour) told us what the parameters were, i.e. defined a space of probability distributions to explore.

Probabilities on minimalist grammars



Probabilities on minimalist grammars



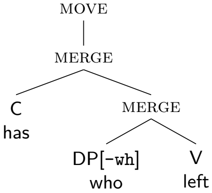
(Things are more complicated because the applicability of a particular rule can't be determined by looking at the directly neighbouring rules in the derivation; cf. n-grams vs. HMMs)

(Hunter and Dyer 2013)

$$P = \frac{\exp(\lambda_{\text{MOVE}} + \lambda_{\texttt{wh}})}{\exp(\lambda_{\text{MOVE}} + \lambda_{\texttt{wh}}) + \exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{v}})} \times \frac{\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{v}})}{\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{v}})} \times \frac{\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{d}})}{\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{d}}) + \exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{c}})}$$

$$\times \frac{\exp(\lambda_{\text{left}})}{\exp(\lambda_{\text{left}}) + \exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{d}})} \times \frac{\exp(\lambda_{\text{has}})}{\exp(\lambda_{\text{has}}) + \exp(\lambda_{\text{will}})} \times \frac{\exp(\lambda_{\text{who}})}{\exp(\lambda_{\text{who}}) + \exp(\lambda_{\text{what}})}$$

Probabilities on minimalist grammars



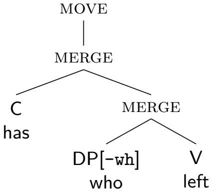
(Things are more complicated because the applicability of a particular rule can't be determined by looking at the directly neighbouring rules in the derivation; cf. n-grams vs. HMMs)

(Hunter and Dyer 2013)

$$P = \frac{\exp(\lambda_{\text{MOVE}} + \lambda_{\texttt{wh}})}{\exp(\lambda_{\text{MOVE}} + \lambda_{\texttt{wh}}) + \exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{v}})} \times \frac{\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{v}})}{\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{v}})} \times \frac{\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{d}})}{\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{d}}) + \exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{c}})}$$
$$\times \frac{\exp(\lambda_{\text{left}})}{\exp(\lambda_{\text{left}}) + \exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{d}})} \times \frac{\exp(\lambda_{\text{has}})}{\exp(\lambda_{\text{has}}) + \exp(\lambda_{\text{will}})} \times \frac{\exp(\lambda_{\text{who}})}{\exp(\lambda_{\text{who}}) + \exp(\lambda_{\text{what}})}$$

Training question: What values of $\lambda_{\text{MERGE}}, \lambda_{\text{MOVE}}, \lambda_{\texttt{wh}}, \lambda_{\texttt{d}}, \ldots$ maximize the likelihood of the training data $P(D|\boldsymbol{\lambda})$? (think blue and red)

Probabilities on minimalist grammars



(Things are more complicated because the applicability of a particular rule can't be determined by looking at the directly neighbouring rules in the derivation; cf. n-grams vs. HMMs)

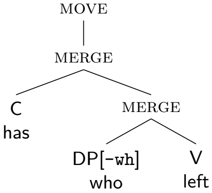
(Hunter and Dyer 2013)

$$P = \frac{\exp(\lambda_{\text{MOVE}} + \lambda_{\texttt{wh}})}{\exp(\lambda_{\text{MOVE}} + \lambda_{\texttt{wh}}) + \exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{v}})} \times \frac{\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{v}})}{\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{v}})} \times \frac{\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{d}})}{\exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{d}}) + \exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{c}})}$$
$$\times \frac{\exp(\lambda_{\text{left}})}{\exp(\lambda_{\text{left}}) + \exp(\lambda_{\text{MERGE}} + \lambda_{\texttt{d}})} \times \frac{\exp(\lambda_{\text{has}})}{\exp(\lambda_{\text{has}}) + \exp(\lambda_{\text{will}})} \times \frac{\exp(\lambda_{\text{who}})}{\exp(\lambda_{\text{who}}) + \exp(\lambda_{\text{what}})}$$

Training question: What values of $\lambda_{\text{MERGE}}, \lambda_{\text{MOVE}}, \lambda_{\texttt{wh}}, \lambda_{\texttt{d}}, \ldots$ maximize the likelihood of the training data $P(D|\boldsymbol{\lambda})$? (think blue and red)

And with IMGs, things will be different: $\lambda_{\text{MRG}}, \lambda_{\text{INSERT}}, \lambda_{\texttt{wh}}, \lambda_{\texttt{d}}, \ldots$ (think green and yellow)

Probabilities on minimalist grammars



(Things are more complicated because the applicability of a particular rule can't be determined by looking at the directly neighbouring rules in the derivation; cf. n-grams vs. HMMs)

(Hunter and Dyer 2013)

$$P = \frac{\exp(\lambda_{\text{MOVE}} + \lambda_{\mathtt{wh}})}{\exp(\lambda_{\text{MOVE}} + \lambda_{\mathtt{wh}}) + \exp(\lambda_{\text{MERGE}} + \lambda_{\mathtt{v}})} \times \frac{\exp(\lambda_{\text{MERGE}} + \lambda_{\mathtt{v}})}{\exp(\lambda_{\text{MERGE}} + \lambda_{\mathtt{v}})} \times \frac{\exp(\lambda_{\text{MERGE}} + \lambda_{\mathtt{d}})}{\exp(\lambda_{\text{MERGE}} + \lambda_{\mathtt{d}}) + \exp(\lambda_{\text{MERGE}} + \lambda_{\mathtt{c}})}$$
$$\times \frac{\exp(\lambda_{\text{left}})}{\exp(\lambda_{\text{left}}) + \exp(\lambda_{\text{MERGE}} + \lambda_{\mathtt{d}})} \times \frac{\exp(\lambda_{\text{has}})}{\exp(\lambda_{\text{has}}) + \exp(\lambda_{\text{will}})} \times \frac{\exp(\lambda_{\text{who}})}{\exp(\lambda_{\text{who}}) + \exp(\lambda_{\text{what}})}$$

Training question: What values of $\lambda_{\text{MERGE}}, \lambda_{\text{MOVE}}, \lambda_{\mathtt{wh}}, \lambda_{\mathtt{d}}, \ldots$ maximize the likelihood of the training data $P(D|\boldsymbol{\lambda})$? (think blue and red)

And with IMGs, things will be different: $\lambda_{\text{MRG}}, \lambda_{\text{INSERT}}, \lambda_{\mathtt{wh}}, \lambda_{\mathtt{d}}, \ldots$ (think green and yellow)

So each system has a different range of probability distributions to explore.

A toy minimalist lexicon

ϵ :: `=t c`	boys :: `=x =det subj`
ϵ :: `=t +wh c`	ϵ :: `x`
will :: `=v =subj t`	some :: `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
some boys will shave
some boys will shave themselves

Some details:

- Subject is base-generated in SpecTP; no movement for Case
- Transitive and intransitive versions of 'shave'
- 'some' is a determiner that optionally combines with 'boys' to make a subject
  - Dummy feature `x` to fill complement of 'boys' so that 'some' 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

Distinct ranges of probability distributions

Take a single training corpus . . .

Possible sentences	Training corpus frequency
boys will shave	10
boys will shave themselves	2
who will shave	3
who will shave themselves	1
some boys will shave	5
some boys will shave themselves	0

Distinct ranges of probability distributions

Take a single training corpus . . .

Possible sentences	Training corpus frequency
boys will shave	10
boys will shave themselves	2
who will shave	3
who will shave themselves	1
some boys will shave	5
some boys will shave themselves	0

Separately ask:

- what values for $\lambda_{\text{MERGE}}, \lambda_{\text{MOVE}}, \lambda_{\mathtt{d}}, \lambda_{\mathtt{wh}}, \dots$ best fit this training data?
- what values for $\lambda_{\text{MRG}}, \lambda_{\text{INSERT}}, \lambda_{\mathtt{d}}, \lambda_{\mathtt{wh}}, \dots$ best fit this training data?

And what do the two results say about the common set of sentences?

Distinct ranges of probability distributions

Take a single training corpus . . .

Possible sentences	Training corpus frequency
boys will shave	10
boys will shave themselves	2
who will shave	3
who will shave themselves	1
some boys will shave	5
some boys will shave themselves	0

Separately ask:

- what values for $\lambda_{\text{MERGE}}, \lambda_{\text{MOVE}}, \lambda_{\mathtt{d}}, \lambda_{\mathtt{wh}}, \ldots$ best fit this training data?
- what values for $\lambda_{\text{MRG}}, \lambda_{\text{INSERT}}, \lambda_{\mathtt{d}}, \lambda_{\mathtt{wh}}, \ldots$ best fit this training data?

And what do the two results say about the common set of sentences?

MG, i.e. MERGE and MOVE	
0.35478	boys will shave
0.35478	some boys will shave
0.14801	who will shave
0.05022	boys will shave themselves
0.05022	some boys will shave themselves
0.04199	who will shave themselves

IMG, i.e. MRG and INSERT	
0.35721	boys will shave
0.35721	some boys will shave
0.095	who will shave
0.095	who will shave themselves
0.04779	boys will shave themselves
0.04779	some boys will shave themselves

Distinct ranges of probability distributions

Take a single training corpus ...

Possible sentences	Training corpus frequency
boys will shave	10
boys will shave themselves	2
who will shave	3
who will shave themselves	1
some boys will shave	5
some boys will shave themselves	0

Separately ask:

- what values for $\lambda_{\text{MERGE}}, \lambda_{\text{MOVE}}, \lambda_{\texttt{d}}, \lambda_{\texttt{wh}}, \dots$ best fit this training data?
- what values for $\lambda_{\text{MRG}}, \lambda_{\text{INSERT}}, \lambda_{\texttt{d}}, \lambda_{\texttt{wh}}, \dots$ best fit this training data?

And what do the two results say about the common set of sentences?

MG, i.e. MERGE and MOVE	
0.35478	boys will shave
0.35478	some boys will shave
0.14801	who will shave
0.05022	boys will shave themselves
0.05022	some boys will shave themselves
0.04199	who will shave themselves

IMG, i.e. MRG and INSERT	
0.35721	boys will shave
0.35721	some boys will shave
0.095	who will shave
0.095	who will shave themselves
0.04779	boys will shave themselves
0.04779	some boys will shave themselves

Choice points in the MG-derived MCFG

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

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

Choice points in the IMG-derived MCFG

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

Non-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 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 MG

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

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

$$P(\text{wh is reflexivized}) = \frac{\exp(\lambda_{\text{MOVE}})}{\exp(\lambda_{\text{MERGE}}) + \exp(\lambda_{\text{MOVE}})} = 0.2213$$
$$P(\text{wh not reflexivized}) = \frac{\exp(\lambda_{\text{MERGE}})}{\exp(\lambda_{\text{MERGE}}) + \exp(\lambda_{\text{MOVE}})} = 0.7787$$

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

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

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

Learned weights on the IMG

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

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

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

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

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

$$P(\text{who will shave}) = 0.5 \times 0.1905 = 0.095$$
$$P(\text{boys will shave themselves}) = 0.5 \times 0.8095 \times 0.1176 = 0.048$$

Surprisal predictions

Grammar: MG, i.e. MERGE and MOVE
Sentence: ‘who will shave themselves’

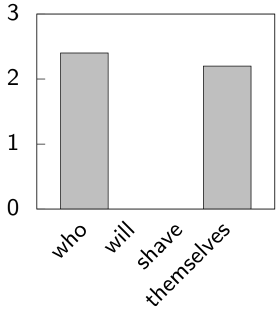
MG, i.e. MERGE and MOVE	
0.35478	boys will shave
0.35478	some boys will shave
0.14801	who will shave
0.05022	boys will shave themselves
0.05022	some boys will shave themselves
0.04199	who will shave themselves

Surprisal predictions

Grammar: MG, i.e. MERGE and MOVE
Sentence: 'who will shave themselves'

MG, i.e. MERGE and MOVE	
0.35478	boys will shave
0.35478	some boys will shave
0.14801	who will shave
0.05022	boys will shave themselves
0.05022	some boys will shave themselves
0.04199	who will shave themselves

$$\begin{aligned}
\text{surprisal at 'who'} &= -\log P(W_1 = \text{who}) \\
&= -\log(0.15 + 0.04) \\
&= -\log 0.19 \\
&= 2.4 \\
\text{surprisal at 'themselves'} &= -\log P(W_4 = \text{themselves} \mid W_1 = \text{who}, \ldots) \\
&= -\log \frac{0.04}{0.15 + 0.04} \\
&= -\log 0.21 \\
&= 2.2
\end{aligned}$$



Surprisal predictions

Grammar: IMG, i.e. MRG and INSERT
Sentence: ‘who will shave themselves’

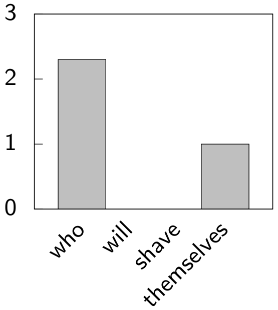
IMG, i.e. MRG and INSERT	
0.35721	boys will shave
0.35721	some boys will shave
0.095	who will shave
0.095	who will shave themselves
0.04779	boys will shave themselves
0.04779	some boys will shave themselves

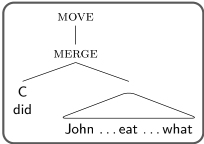
Surprisal predictions

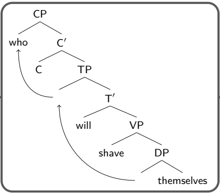
Grammar: IMG, i.e. MRG and INSERT
Sentence: 'who will shave themselves'

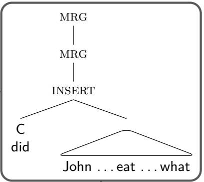
IMG, i.e. MRG and INSERT	
0.35721	boys will shave
0.35721	some boys will shave
0.095	who will shave
0.095	who will shave themselves
0.04779	boys will shave themselves
0.04779	some boys will shave themselves

$$\begin{aligned}
\text{surprisal at 'who'} &= -\log P(W_1 = \text{who}) \\
&= -\log(0.10 + 0.10) \\
&= -\log 0.2 \\
&= 2.3 \\
\text{surprisal at 'themselves'} &= -\log P(W_4 = \text{themselves} \mid W_1 = \text{who}, \ldots) \\
&= -\log \frac{0.10}{0.10 + 0.10} \\
&= -\log 0.5 \\
&= 1
\end{aligned}$$





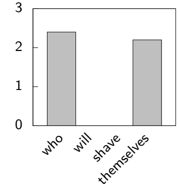


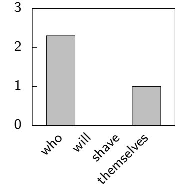


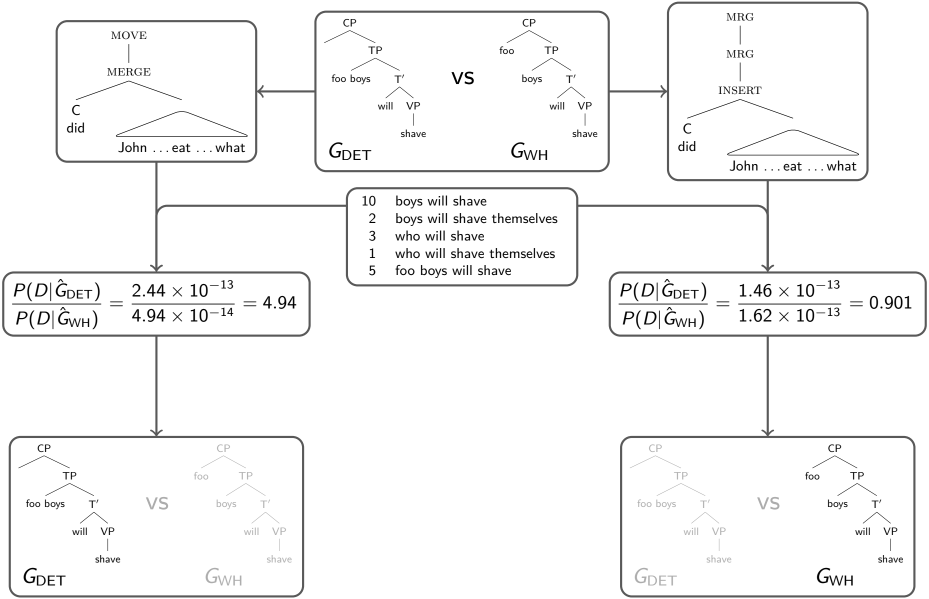
10	boys will shave
2	boys will shave themselves
3	who will shave
1	who will shave themselves
5	some boys will shave

MG, i.e. MERGE and MOVE	
0.35478	boys will shave
0.35478	some boys will shave
0.14801	who will shave
0.05022	boys will shave themselves
0.05022	some boys will shave themselves
0.04199	who will shave themselves

IMG, i.e. MRG and INSERT	
0.35721	boys will shave
0.35721	some boys will shave
0.095	who will shave
0.095	who will shave themselves
0.04779	boys will shave themselves
0.04779	some boys will shave themselves







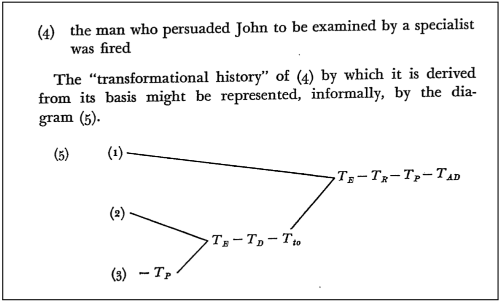
MOVE
MERGE
C
did
John . . . eat . . . what
CP
TP
foo boys
T′
will
VP
shave
G_{DET}
VS
CP
foo
TP
boys
T′
will
VP
shave
G_{WH}
MRG
MRG
INSERT
C
did
John . . . eat . . . what
10 boys will shave
2 boys will shave themselves
3 who will shave
1 who will shave themselves
5 foo boys will shave
$\frac{P(D|\hat{G}_{\text{DET}})}{P(D|\hat{G}_{\text{WH}})} = \frac{2.44 \times 10^{-13}}{4.94 \times 10^{-14}} = 4.94$
$\frac{P(D|\hat{G}_{\text{DET}})}{P(D|\hat{G}_{\text{WH}})} = \frac{1.46 \times 10^{-13}}{1.62 \times 10^{-13}} = 0.901$
CP
TP
foo boys
T′
will
VP
shave
G_{DET}
VS
CP
foo
TP
boys
T′
will
VP
shave
G_{WH}

Outline

Back in the good old (heavily derivational) days

[The perceptual model] will utilize the full resources of the transformational grammar to provide a structural description, consisting of a set of P-markers and a transformational history

(Miller and Chomsky 1963: p.480)


(4) the man who persuaded John to be examined by a specialist was fired

The "transformational history" of (4) by which it is derived from its basis might be represented, informally, by the diagram (5).

Back in the good old (heavily derivational) days

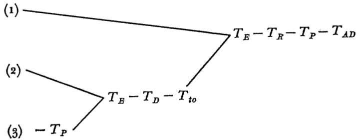
[The perceptual model] will utilize the full resources of the transformational grammar to provide a structural description, consisting of a set of P-markers and a transformational history

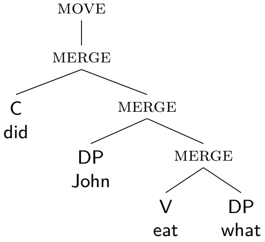
(Miller and Chomsky 1963: p.480)

(4) the man who persuaded John to be examined by a specialist was fired

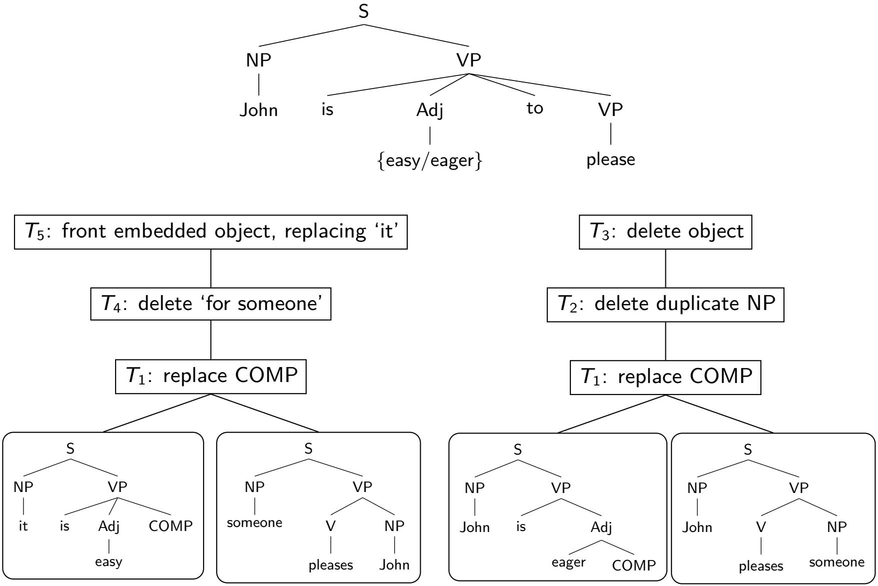
The "transformational history" of (4) by which it is derived from its basis might be represented, informally, by the diagram (5).

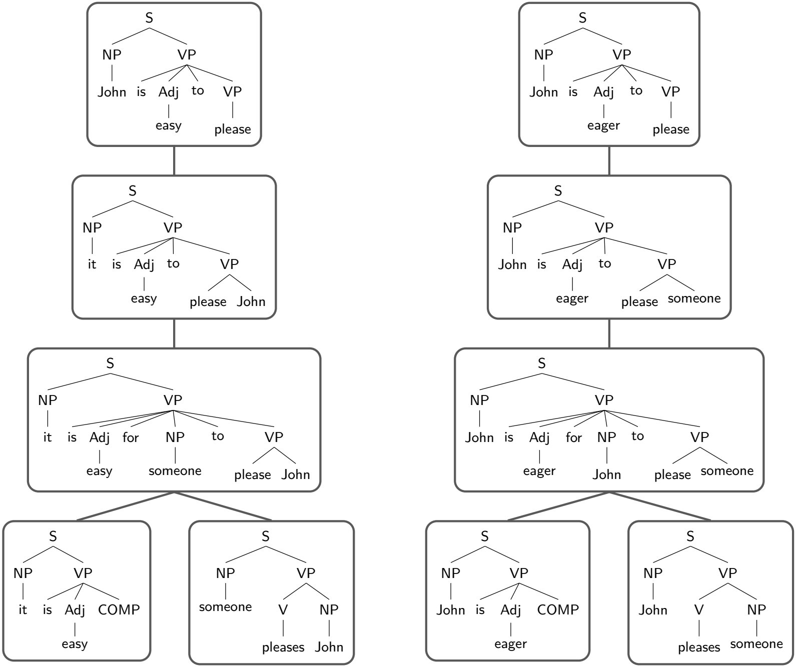
(5)





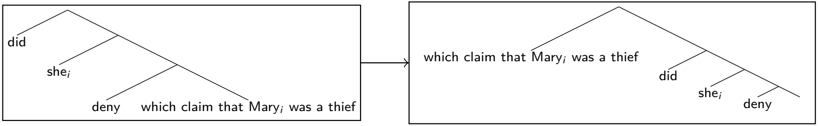
Full derivational histories




S
NP VP
John is Adj to VP
easy please
S
NP VP
it is Adj to VP
easy please John
S
NP VP
it is Adj for NP to VP
easy someone please John
S
NP VP
it is Adj COMP
easy
S
NP VP
someone V NP
pleases John
S
NP VP
John is Adj to VP
eager please
S
NP VP
John is Adj to VP
eager please someone
S
NP VP
John is Adj for NP to VP
eager John please someone
S
NP VP
John is Adj COMP
eager
S
NP VP
John V NP
pleases someone

Full derivational histories

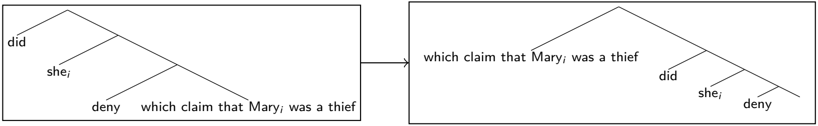
(1) * Which claim [that Mary$_i$ was a thief] did she$_i$ deny?



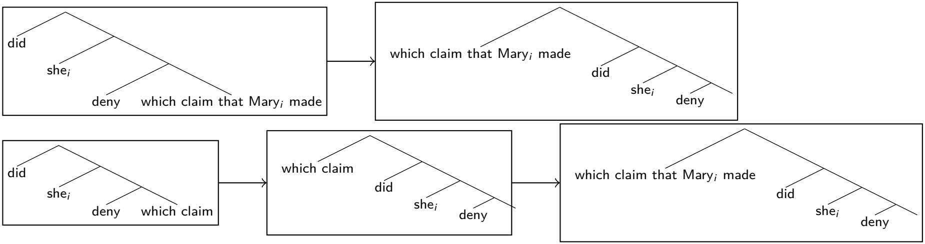
(2) Which claim [that Mary$_i$ made] did she$_i$ deny?

Full derivational histories

(1) * Which claim [that Mary$_i$ was a thief] did she$_i$ deny?

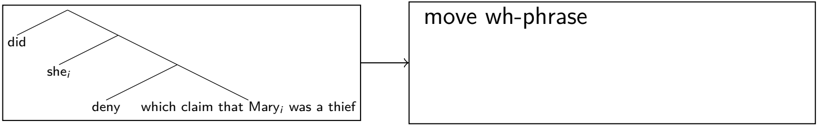


(2) Which claim [that Mary$_i$ made] did she$_i$ deny?

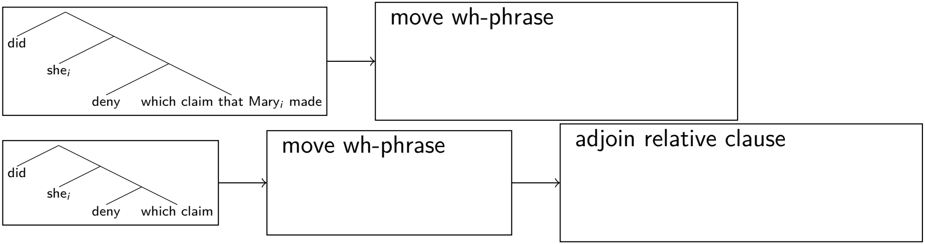


Full derivational histories

(1) * Which claim [that Mary$_i$ was a thief] did she$_i$ deny?



(2) Which claim [that Mary$_i$ made] did she$_i$ deny?



Conclusion and open issues

We can formulate a theory where

- the choice of derivational operations has empirically-testable consequences (so we are not extensionalists), and
- this does not happen by taking derivational operations to be real-time operations (because we are not literalists).

Conclusion and open issues

We can formulate a theory where

- the choice of derivational operations has empirically-testable consequences (so we are not extensionalists), and
- this does not happen by taking derivational operations to be real-time operations (because we are not literalists).

Unanswered question: "So **what are** the real-time operations?"

- For complementary proposals on this see Stabler (2013), Kobele et al. (2013), Graf et al. (2015), Hunter (forthcoming)
- . . . but all of these take the form of procedures for identifying a derivation tree/T-marker.
- Distinct from the question of what the chunks to be (somehow) chained together are.

References I

Graf, T., Fodor, B., Monette, J., Rachiele, G., Warren, A., and Zhang, C. (2015). A refined notion of memory usage for minimalist parsing. In *Proceedings of the 14th Meeting on the Mathematics of Language (MoL 2015)*, pages 1–14. 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. (forthcoming). Left-corner parsing of minimalist grammars. In Berwick, B. and Stabler, E., editors, *Minimalist Parsing*. Oxford University Press.

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

Keenan, E. L. and Stabler, E. P. (2003). *Bare Grammar*. CSLI Publications, Stanford, CA.

Kobele, G. M., Gerth, S., and Hale, J. (2013). Memory resource allocation in top-down minimalist parsing. In Morrill, G. and Nederhof, M.-J., editors, *Formal Grammar 2012/2013*, volume 8036 of *Lecture Notes in Computer Science*, pages 32–51. Springer.

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.

References II

Phillips, C. and Lewis, S. (2013). Derivational order in syntax: evidence and architectural consequences. *Studies in Linguistics*, 6:11–47.

Stabler, E. (2013). Two models of minimalist, incremental syntactic analysis. *Topics in Cognitive Science*, 5(3):611–633.

Stabler, E. P. (1997). Derivational minimalism. In Retoré, C., editor, *Logical Aspects of Computational Linguistics*, volume 1328 of *LNCS*, pages 68–95, Berlin Heidelberg. Springer.

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.