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

1. What are grammars?

2. Derivationally distinct implementations of merge

3. Telling them apart

4. Historical perspective
Outline

1. What are grammars?
2. Derivationally distinct implementations of merge
3. Telling them apart
4. Historical perspective
What is a grammar?

**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
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.
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?
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
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
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 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 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.
Three views of derivations: literalist, formalist, extensionalist

These correspond to different views about which linking hypotheses expose 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.
Three views of derivations: literalist, formalist, extensionalist

These correspond to different views about which linking hypotheses expose 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.
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 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’.
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’.

Phillips and Lewis (2013)
What are grammars?

Contemporary syntactic derivations

How does a speaker recognize this as a well-formed expression?
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 \( \text{TP} \) is a well-formed expression by

- knowing that \( \text{DP} \) is well-formed if \( \text{DP} \) and \( \text{T'} \) are well-formed

- recognizing that \( \text{DP} \) is well-formed

- recognizing that \( \text{T'} \) is well-formed
A speaker recognizes that \( \text{TP} \) is a well-formed expression by

- knowing that \( \text{DP} \) is well-formed if \( \text{T} \) and \( \text{T'} \) are well-formed

- recognizing that \( \text{DP} \) is well-formed

- recognizing that \( \text{T} \) is well-formed

A speaker recognizes that \( \text{TP} \) is a well-formed expression by

- knowing that \( \text{DP} \) is well-formed if \( \text{D} \) and \( \text{N} \) are well-formed

- knowing that \( \text{D} \) is well-formed

- knowing that \( \text{N} \) is well-formed
A speaker recognizes that \( \text{DP} \) \( \text{N boy} \) \( \text{the} \) is a well-formed expression by knowing that 
- \( \text{DP} \) \( \text{N} \) \( \text{D} \) \( \text{the} \) \( \text{boy} \) is well-formed if \( \text{D} \) and \( \text{N} \) are well-formed
- \( \text{DP} \) \( \text{N} \) \( \text{D} \) \( \text{the} \) \( \text{boy} \) is well-formed
- \( \text{DP} \) \( \text{N} \) \( \text{D} \) \( \text{the} \) \( \text{boy} \) is well-formed

A speaker recognizes that \( \text{T}\)' \( \text{V Prun} \) \( \text{T will} \) \( \text{run} \) is a well-formed expression by knowing that 
- \( \text{T}\)' \( \text{V Prun} \) \( \text{T will} \) \( \text{run} \) is well-formed if \( \text{T} \) and \( \text{VP} \) are well-formed
- \( \text{T}\)' \( \text{V Prun} \) \( \text{T will} \) \( \text{run} \) is well-formed
- \( \text{T}\)' \( \text{V Prun} \) \( \text{T will} \) \( \text{run} \) is well-formed
A speaker recognizes that \( \text{the boy} \) is a well-formed expression by

- knowing that \( \text{the} \) is well-formed if \( \text{boy} \) and \( \text{the} \) are well-formed (merge)
- recognizing that \( \text{boy} \) is well-formed
- recognizing that \( \text{will} \) is well-formed

A speaker recognizes that \( \text{the boy will run} \) is a well-formed expression by

- knowing that \( \text{the} \) is well-formed if \( \text{boy} \) and \( \text{the} \) are well-formed (merge)
- knowing that \( \text{will} \) is well-formed
- knowing that \( \text{run} \) is well-formed

A speaker recognizes that \( \text{the boy will run} \) is a well-formed expression by

- knowing that \( \text{will} \) is well-formed if \( \text{run} \) and \( \text{will} \) are well-formed (merge)
- knowing that \( \text{run} \) is well-formed
A speaker recognizes that \( TP \) is a well-formed expression by

- knowing that \( TP \) is well-formed if \( DP \) and \( T' \) are well-formed (MERGE)
  
- recognizing that \( DP \) is well-formed

- recognizing that \( T' \) is well-formed

A speaker recognizes that \( DP \) is a well-formed expression by

- knowing that \( DP \) is well-formed if \( D \) and \( N \) are well-formed (MERGE)
  
- knowing that \( D \) is well-formed

- knowing that \( N \) is well-formed

A speaker recognizes that \( T' \) is a well-formed expression by

- knowing that \( T' \) is well-formed if \( T \) and \( VP \) are well-formed (MERGE)
  
- knowing that \( T \) is well-formed

- knowing that \( VP \) 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
What are grammars? Derivationally distinct implementations of merge

A speaker recognizes that \( \text{CP} \rightarrow \text{DP}_i \text{C}^t \text{VP} \) is a well-formed expression by

- knowing that \( \text{CP} \rightarrow \text{DP}_i \text{C}^t \text{VP} \) is well-formed if \( \text{C}^t \text{VP} \rightarrow \text{VP} \) is well-formed

- recognizing that \( \text{CP} \rightarrow \text{DP}_i \text{C}^t \text{VP} \) is well-formed

A speaker recognizes that \( \text{C} \rightarrow \text{VP} \) is a well-formed expression by

- knowing that \( \text{C} \rightarrow \text{VP} \) is well-formed if \( \text{C} \) and \( \text{VP} \) are well-formed

- knowing that \( \text{C} \rightarrow \text{VP} \) is well-formed

- recognizing that \( \text{C} \rightarrow \text{VP} \) 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

- knowing that \( C \) is well-formed if \( C' \) is well-formed (MOVE)

- recognizing that \( C \) is well-formed

A speaker recognizes that

- knowing that \( C' \) is well-formed if \( C \) and \( VP \) are well-formed (MERGE)

- knowing that \( C \) is well-formed

- recognizing that \( VP \) 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:

```
MOVE
  |   MERGE
  C did

MERGE
  DP John
  V eat

MERGE
  V t_i eat
  DP[-wh] what
```
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

So it should pattern with (5, 4, 6).

Both have probability: $P(4) \cdot P(5) \cdot P(6)$

**Hypothesis #2:**

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

So it should pattern with (4, 6, 5).

Both have probability: $P(4) \cdot P(5) \cdot P(6)$

Same possibilities: both recognize the set {1, 2, 3, 4, 5, 6}

Different probabilities: distinct ranges of probability distributions over this set

More generally: distinct similarity relations over this set
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

So it should pattern with (5, 4, 6).

Both have probability:

\[ P(4) \cdot P(5) \cdot P(6) \]

**Hypothesis #2:**

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

So it should pattern with (4, 6, 5).

Both have probability:

\[ P(4) \cdot P(5) \cdot P(6) \]

Same possibilities: both recognize the set \{1, 2, 3, 4, 5, 6\}

Different probabilities: distinct ranges of probability distributions over this set

More generally: distinct similarity relations over this set
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

Same possibilities: both recognize the set \(\{1, 2, 3, 4, 5, 6\}\)

Different probabilities: distinct ranges of probability distributions over this set

More generally: distinct similarity relations over this set
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: \(P(4) \cdot P(5) \cdot P(6)\)

**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: \(P(4) \cdot P(5) \cdot P(6)\)
What are grammars? Derivationally distinct implementations of merge Telling them apart Historical perspective

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: \(P(4) \cdot P(5) \cdot P(6)\)

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

**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: \(P(4) \cdot P(5) \cdot P(6)\)
Outline

1. What are grammars?

2. Derivationally distinct implementations of merge

3. Telling them apart

4. Historical perspective
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
A speaker recognizes that this is a well-formed expression by

- knowing that $\text{CP} \rightarrow \text{C}' \rightarrow \text{V}' \rightarrow \text{t}_i \rightarrow \text{VP}$ is well-formed if $\text{DP} \rightarrow \text{C}' \rightarrow \text{t}_i \rightarrow \text{VP}$ is well-formed (MRG)

- recognizing that $\text{DP} \rightarrow \text{C}' \rightarrow \text{V}' \rightarrow \text{t}_i \rightarrow \text{eat}$ is well-formed
IMG derivations

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

- knowing that \( \text{DP}_i \) is well-formed if \( \text{CP} \) is well-formed (MRG)

- recognizing that \( \text{VP} \) is well-formed

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

- knowing that \( \text{DP}[-\text{wh}]_i \) is well-formed if \( \text{C} \) is well-formed (MRG)

- recognizing that \( \text{VP} \) is well-formed
**IMG derivations**

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

- knowing that \( \text{DP}_i \) is well-formed if \( \text{VP}[\text{-wh}]_i \) is well-formed (MRG)

- recognizing that \( \text{VP}[\text{-wh}]_i \) is well-formed

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

- knowing that \( \text{DP}[\text{-wh}]_i \) is well-formed if \( \text{VP}[\text{-wh}]_i \) is well-formed (MRG)

- recognizing that \( \text{C} \) is well-formed
A speaker recognizes that

\[
\begin{array}{c}
\text{C did} \\
\text{DP[-wh], what}
\end{array}
\]

is well-formed by

- knowing that

\[
\begin{array}{c}
\text{C did} \\
\text{DP[-wh], ti}
\end{array}
\]

is well-formed if

\[
\begin{array}{c}
\text{C did} \\
\text{DP[-wh], ti}
\end{array}
\]

are well-formed (\text{INSERT})

- knowing that

\[
\begin{array}{c}
\text{C did} \\
\text{DP[-wh], ti}
\end{array}
\]

is well-formed

- recognizing that

\[
\begin{array}{c}
\text{DP[-wh], what} \\
\text{DP, John} \\
\text{V, ti}
\end{array}
\]

is well-formed
A speaker recognizes that \( C \text{ did} \) is well-formed by

- knowing that \( C \text{ did} \) is well-formed
- recognizing that \( C \text{ did} \) is well-formed

and

- knowing that \( C \text{ did} \) is well-formed if \( C \) and \( \text{DP[-wh], i} \) are well-formed (INSERT)
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

MRG

MRG

INSERT

C
did

John ... eat ... what
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**

```
  MOVE
   \-- MERGE
       \-- C
           \-- dbo
               \-- C
                   \-- VP
                       \-- V
                           \-- V
                               \-- t_i
                       \-- V'
                           \-- t_i
               \-- DP
                   \-- John
               \-- DP
                   \-- DP
                       \-- V
                           \-- V
                               \-- what
               \-- C
                   \-- did
               \-- V
                   \-- eat
```

IMG hypothesis: **MRG** and **INSERT**

```
  MRG
   \-- MRG
       \-- INSERT
           \-- C
               \-- did
           \-- V
               \-- eat
           \-- DP
               \-- what
           \-- DP
               \-- DP
                   \-- V
                       \-- V
                           \-- what
```

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

1. What are grammars?
2. Derivationally distinct implementations of merge
3. Telling them apart
4. Historical perspective
Probabilities on a CFG

S → NP VP
NP → John
NP → he
NP → D N
D → the
D → a
N → dog
N → cat
VP → V NP
VP → V
V → chased
V → ate
Probabilities on a CFG

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

1.0  \( S \rightarrow \text{NP} \ \text{VP} \)
0.3  \( \text{NP} \rightarrow \text{John} \)
0.2  \( \text{NP} \rightarrow \text{he} \)
0.5  \( \text{NP} \rightarrow \text{D} \ \text{N} \)
0.3  \( \text{D} \rightarrow \text{the} \)
0.7  \( \text{D} \rightarrow \text{a} \)
0.6  \( \text{N} \rightarrow \text{dog} \)
0.4  \( \text{N} \rightarrow \text{cat} \)
0.8  \( \text{VP} \rightarrow \text{V} \ \text{NP} \)
0.2  \( \text{VP} \rightarrow \text{V} \)
0.9  \( \text{V} \rightarrow \text{chased} \)
0.1  \( \text{V} \rightarrow \text{ate} \)
Probabilities on a CFG

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

- \lambda_1 \quad S \rightarrow NP \ VP
- \lambda_2 \quad NP \rightarrow John
- \lambda_3 \quad NP \rightarrow he
- \lambda_4 \quad NP \rightarrow D \ N
- \lambda_5 \quad D \rightarrow the
- \lambda_6 \quad D \rightarrow a
- \lambda_7 \quad N \rightarrow dog
- \lambda_8 \quad N \rightarrow cat
- \lambda_9 \quad VP \rightarrow V \ NP
- \lambda_{10} \quad VP \rightarrow V
- \lambda_{11} \quad V \rightarrow chased
- \lambda_{12} \quad V \rightarrow ate
Probabilities on a CFG

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

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

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

\[
P(D|\lambda) = \exp(\lambda_{\text{move}} + \lambda_{\text{wh}}) + \exp(\lambda_{\text{merge}} + \lambda_{\text{v}}) \times \exp(\lambda_{\text{merge}} + \lambda_{\text{d}}) + \exp(\lambda_{\text{merge}} + \lambda_{\text{c}}) \times \exp(\lambda_{\text{left}}) + \exp(\lambda_{\text{merge}} + \lambda_{\text{d}}) \times \exp(\lambda_{\text{has}}) + \exp(\lambda_{\text{will}}) \times \exp(\lambda_{\text{who}}) + \exp(\lambda_{\text{what}})
\]

Training question: What values of \(\lambda_{\text{merge}}, \lambda_{\text{move}}, \lambda_{\text{wh}}, \lambda_{\text{d}}, \ldots\) maximize the likelihood of the training data? (think blue and red)

And with IMGs, things will be different: \(\lambda_{\text{mrg}}, \lambda_{\text{insert}}, \lambda_{\text{wh}}, \lambda_{\text{d}}, \ldots\) (think green and yellow)

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

\[
P = \frac{\exp(\lambda_{\text{MOVE}} + \lambda_{\text{wh}})}{\exp(\lambda_{\text{MOVE}} + \lambda_{\text{wh}}) + \exp(\lambda_{\text{MERGE}} + \lambda_{V})} \times \frac{\exp(\lambda_{\text{MERGE}} + \lambda_{V})}{\exp(\lambda_{\text{MERGE}} + \lambda_{V})} \times \frac{\exp(\lambda_{\text{MERGE}} + \lambda_{d})}{\exp(\lambda_{\text{MERGE}} + \lambda_{d}) + \exp(\lambda_{\text{MERGE}} + \lambda_{c})} \times \frac{\exp(\lambda_{\text{left}})}{\exp(\lambda_{\text{left}}) + \exp(\lambda_{\text{MERGE}} + \lambda_{d})} \times \frac{\exp(\lambda_{\text{has}})}{\exp(\lambda_{\text{has}}) + \exp(\lambda_{\text{will}})} \times \frac{\exp(\lambda_{\text{who}})}{\exp(\lambda_{\text{who}}) + \exp(\lambda_{\text{what}})}
\]
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)

\[
P = \frac{\exp(\lambda_{\text{MOVE}} + \lambda_{\text{wh}})}{\exp(\lambda_{\text{MOVE}} + \lambda_{\text{wh}}) + \exp(\lambda_{\text{MERGE}} + \lambda_{V})} \times \frac{\exp(\lambda_{\text{MERGE}} + \lambda_{V})}{\exp(\lambda_{\text{MERGE}} + \lambda_{d}) + \exp(\lambda_{\text{MERGE}} + \lambda_{c})} \times \frac{\exp(\lambda_{\text{MERGE}} + \lambda_{d})}{\exp(\lambda_{\text{MERGE}} + \lambda_{d}) + \exp(\lambda_{\text{MERGE}} + \lambda_{c})} \\
\times \frac{\exp(\lambda_{\text{left}})}{\exp(\lambda_{\text{left}}) + \exp(\lambda_{\text{MERGE}} + \lambda_{d})} \times \frac{\exp(\lambda_{\text{has}})}{\exp(\lambda_{\text{has}}) + \exp(\lambda_{\text{will}})} \times \frac{\exp(\lambda_{\text{who}})}{\exp(\lambda_{\text{who}}) + \exp(\lambda_{\text{what}})}
\]

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

$P = \frac{\exp(\lambda_{\text{move}} + \lambda_{\text{wh}})}{\exp(\lambda_{\text{move}} + \lambda_{\text{wh}}) + \exp(\lambda_{\text{merge}} + \lambda_{v})} \times \frac{\exp(\lambda_{\text{merge}} + \lambda_{v})}{\exp(\lambda_{\text{merge}} + \lambda_{v}) + \exp(\lambda_{\text{merge}} + \lambda_{d})} \times \frac{\exp(\lambda_{\text{merge}} + \lambda_{d})}{\exp(\lambda_{\text{merge}} + \lambda_{d}) + \exp(\lambda_{\text{merge}} + \lambda_{c})} \times \frac{\exp(\lambda_{\text{left}})}{\exp(\lambda_{\text{left}}) + \exp(\lambda_{\text{merge}} + \lambda_{d})} \times \frac{\exp(\lambda_{\text{has}})}{\exp(\lambda_{\text{has}}) + \exp(\lambda_{\text{will}})} \times \frac{\exp(\lambda_{\text{will}})}{\exp(\lambda_{\text{will}}) + \exp(\lambda_{\text{what}})}$

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

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

\[
P = \frac{\exp(\lambda_{\text{MOVE}} + \lambda_{\text{wh}})}{\exp(\lambda_{\text{MOVE}} + \lambda_{\text{wh}}) + \exp(\lambda_{\text{MERGE}} + \lambda_{V})} \times \frac{\exp(\lambda_{\text{MERGE}} + \lambda_{V})}{\exp(\lambda_{\text{MERGE}} + \lambda_{V})} \times \frac{\exp(\lambda_{\text{MERGE}} + \lambda_{d})}{\exp(\lambda_{\text{MERGE}} + \lambda_{d}) + \exp(\lambda_{\text{MERGE}} + \lambda_{c})} \times \frac{\exp(\lambda_{\text{left}})}{\exp(\lambda_{\text{left}}) + \exp(\lambda_{\text{MERGE}} + \lambda_{d})} \times \frac{\exp(\lambda_{\text{has}})}{\exp(\lambda_{\text{has}}) + \exp(\lambda_{\text{will}})} \times \frac{\exp(\lambda_{\text{who}})}{\exp(\lambda_{\text{who}}) + \exp(\lambda_{\text{what}})}
\]

(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)

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

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

So each system has a different range of probability distributions to explore.
A toy minimalist lexicon

\[
\begin{align*}
\epsilon :: &= t \ c \\
\epsilon :: &= t +wh \ c \\
\text{will} :: &= v \ =\text{subj} \ t \\
\text{shave} :: &= v \\
\text{shave} :: &= \text{obj} \ v \\
\text{boys} :: &= \text{subj} \\
\text{who} :: &= \text{subj} \ -wh
\end{align*}
\]

\[
\begin{align*}
\text{boys} :: &= x =\text{det} \ \text{subj} \\
\epsilon :: &= x \\
\text{some} :: &= \text{det} \\
\text{themselves} :: &= \text{ant} \ \text{obj} \\
\epsilon :: &= \text{subj} \ \text{ant} \ -\text{subj} \\
\text{will} :: &= v +\text{subj} \ t
\end{align*}
\]

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 . . .

<table>
<thead>
<tr>
<th>Possible sentences</th>
<th>Training corpus frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>boys will shave</td>
<td>10</td>
</tr>
<tr>
<td>boys will shave themselves</td>
<td>2</td>
</tr>
<tr>
<td>who will shave</td>
<td>3</td>
</tr>
<tr>
<td>who will shave themselves</td>
<td>1</td>
</tr>
<tr>
<td>some boys will shave</td>
<td>5</td>
</tr>
<tr>
<td>some boys will shave themselves</td>
<td>0</td>
</tr>
</tbody>
</table>

Separately ask:
what values for $\lambda_{\text{merge}}$, $\lambda_{\text{move}}$, $\lambda_{\text{d}}$, $\lambda_{\text{wh}}$, ... best fit this training data?

what values for $\lambda_{\text{mrg}}$, $\lambda_{\text{insert}}$, $\lambda_{\text{d}}$, $\lambda_{\text{wh}}$, ... 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 . . .

<table>
<thead>
<tr>
<th>Possible sentences</th>
<th>Training corpus frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>boys will shave</td>
<td>10</td>
</tr>
<tr>
<td>boys will shave themselves</td>
<td>2</td>
</tr>
<tr>
<td>who will shave</td>
<td>3</td>
</tr>
<tr>
<td>who will shave themselves</td>
<td>1</td>
</tr>
<tr>
<td>some boys will shave</td>
<td>5</td>
</tr>
<tr>
<td>some boys will shave themselves</td>
<td>0</td>
</tr>
</tbody>
</table>

Separately ask:

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

<table>
<thead>
<tr>
<th>Possible sentences</th>
<th>Training corpus frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>boys will shave</td>
<td>10</td>
</tr>
<tr>
<td>boys will shave themselves</td>
<td>2</td>
</tr>
<tr>
<td>who will shave</td>
<td>3</td>
</tr>
<tr>
<td>who will shave themselves</td>
<td>1</td>
</tr>
<tr>
<td>some boys will shave</td>
<td>5</td>
</tr>
<tr>
<td>some boys will shave themselves</td>
<td>0</td>
</tr>
</tbody>
</table>

Separately ask:
- what values for $\lambda_{\text{MERGE}}, \lambda_{\text{MOVE}}, \lambda_d, \lambda_{\text{wh}}, \ldots$ best fit this training data?
- what values for $\lambda_{\text{MRG}}, \lambda_{\text{INSERT}}, \lambda_d, \lambda_{\text{wh}}, \ldots$ best fit this training data?

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

<table>
<thead>
<tr>
<th>MG, i.e. MERGE and MOVE</th>
<th>IMG, i.e. MRG and INSERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.35478 boys will shave</td>
<td>0.35721 boys will shave</td>
</tr>
<tr>
<td>0.35478 some boys will shave</td>
<td>0.35721 some boys will shave</td>
</tr>
<tr>
<td>0.14801 who will shave</td>
<td>0.095 who will shave</td>
</tr>
<tr>
<td>0.05022 boys will shave themselves</td>
<td>0.095 who will shave theirselves</td>
</tr>
<tr>
<td>0.05022 some boys will shave theirselves</td>
<td>0.04779 boys will shave theirselves</td>
</tr>
<tr>
<td>0.04199 who will shave theirselves</td>
<td>0.04779 some boys will shave theirselves</td>
</tr>
</tbody>
</table>
What are grammars? Derivationally distinct implementations of merge Telling them apart Historical perspective

Distinct ranges of probability distributions

Take a single training corpus ... 

<table>
<thead>
<tr>
<th>Possible sentences</th>
<th>Training corpus frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>boys will shave</td>
<td>10</td>
</tr>
<tr>
<td>boys will shave themselves</td>
<td>2</td>
</tr>
<tr>
<td>who will shave</td>
<td>3</td>
</tr>
<tr>
<td>who will shave themselves</td>
<td>1</td>
</tr>
<tr>
<td>some boys will shave</td>
<td>5</td>
</tr>
<tr>
<td>some boys will shave themselves</td>
<td>0</td>
</tr>
</tbody>
</table>

Separately ask:
- what values for $\lambda_{\text{MERGE}}, \lambda_{\text{MOVE}}, \lambda_d, \lambda_{\text{wh}}, \ldots$ best fit this training data?
- what values for $\lambda_{\text{Mrg}}, \lambda_{\text{Insert}}, \lambda_d, \lambda_{\text{wh}}, \ldots$ best fit this training data?

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

<table>
<thead>
<tr>
<th>MG, i.e. MERGE and MOVE</th>
<th>IMG, i.e. MRG and INSERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.35478 boys will shave</td>
<td>0.35721 boys will shave</td>
</tr>
<tr>
<td>0.35478 some boys will shave</td>
<td>0.35721 some boys will shave</td>
</tr>
<tr>
<td>0.14801 who will shave</td>
<td>0.095 who will shave</td>
</tr>
<tr>
<td>0.05022 boys will shave themselves</td>
<td>0.095 who will shave themselves</td>
</tr>
<tr>
<td>0.05022 some boys will shave themselves</td>
<td>0.04779 boys will shave themselves</td>
</tr>
<tr>
<td>0.04199 who will shave themselves</td>
<td>0.04779 some boys will shave themselves</td>
</tr>
</tbody>
</table>
## Choice points in the MG-derived MCFG

### Question or not?

<table>
<thead>
<tr>
<th>Rule</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \langle c \rangle_0 \rightarrow \langle =t \ c \rangle_0 \langle t \rangle_0 )</td>
<td>( \exp(\lambda_{\text{merge}} + \lambda_t) )</td>
</tr>
<tr>
<td>( \langle c \rangle_0 \rightarrow \langle +\text{wh} \ c, -\text{wh} \rangle_0 )</td>
<td>( \exp(\lambda_{\text{move}} + \lambda_{\text{wh}}) )</td>
</tr>
</tbody>
</table>

### Non-wh antecedent lexical or complex?

<table>
<thead>
<tr>
<th>Rule</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \langle \text{ant} -\text{subj} \rangle_0 \rightarrow \langle =\text{subj} \ \text{ant} -\text{subj} \rangle_1 \langle \text{subj} \rangle_0 )</td>
<td>( \exp(\lambda_{\text{merge}} + \lambda_{\text{subj}}) )</td>
</tr>
<tr>
<td>( \langle \text{ant} -\text{subj} \rangle_0 \rightarrow \langle =\text{subj} \ \text{ant} -\text{subj} \rangle_1 \langle \text{subj} \rangle_1 )</td>
<td>( \exp(\lambda_{\text{merge}} + \lambda_{\text{subj}}) )</td>
</tr>
</tbody>
</table>

### Non-wh subject merged and complex, merged and lexical, or moved?

<table>
<thead>
<tr>
<th>Rule</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \langle t \rangle_0 \rightarrow \langle =\text{subj} \ t \rangle_0 \langle \text{subj} \rangle_0 )</td>
<td>( \exp(\lambda_{\text{merge}} + \lambda_{\text{subj}}) )</td>
</tr>
<tr>
<td>( \langle t \rangle_0 \rightarrow \langle =\text{subj} \ t \rangle_0 \langle \text{subj} \rangle_1 )</td>
<td>( \exp(\lambda_{\text{merge}} + \lambda_{\text{subj}}) )</td>
</tr>
<tr>
<td>( \langle t \rangle_0 \rightarrow \langle +\text{subj} \ t, -\text{subj} \rangle_0 )</td>
<td>( \exp(\lambda_{\text{move}} + \lambda_{\text{subj}}) )</td>
</tr>
</tbody>
</table>

### Wh-phrase same as subject or separated because of doubling?

<table>
<thead>
<tr>
<th>Rule</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \langle t, -\text{wh} \rangle_0 \rightarrow \langle =\text{subj} \ t \rangle_0 \langle \text{subj} -\text{wh} \rangle_1 )</td>
<td>( \exp(\lambda_{\text{merge}} + \lambda_{\text{subj}}) )</td>
</tr>
<tr>
<td>( \langle t, -\text{wh} \rangle_0 \rightarrow \langle +\text{subj} \ t, -\text{subj}, -\text{wh} \rangle_0 )</td>
<td>( \exp(\lambda_{\text{move}} + \lambda_{\text{subj}}) )</td>
</tr>
</tbody>
</table>
### Choice points in the IMG-derived MCFG

#### Question or not?

<table>
<thead>
<tr>
<th>Context</th>
<th>Transition</th>
<th>Feature</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\langle -c \rangle_0$ → $\langle +t -c, -t \rangle_1$</td>
<td>$\exp(\lambda_{\text{MRG}} + \lambda_t)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\langle -c \rangle_0$ → $\langle +wh -c, -wh \rangle_0$</td>
<td>$\exp(\lambda_{\text{MRG}} + \lambda_{\text{wh}})$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Non-wh antecedent lexical or complex?

<table>
<thead>
<tr>
<th>Context</th>
<th>Transition</th>
<th>Feature</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\langle +\text{subj} -\text{ant} -\text{subj}, -\text{subj} \rangle_0$ → $\langle +\text{subj} -\text{ant} -\text{subj} \rangle_0 \langle -\text{subj} \rangle_0$</td>
<td>$\exp(\lambda_{\text{INSERT}})$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\langle +\text{subj} -\text{ant} -\text{subj}, -\text{subj} \rangle_0$ → $\langle +\text{subj} -\text{ant} -\text{subj} \rangle_0 \langle -\text{subj} \rangle_1$</td>
<td>$\exp(\lambda_{\text{INSERT}})$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Non-wh subject merged and complex, merged and lexical, or moved?

<table>
<thead>
<tr>
<th>Context</th>
<th>Transition</th>
<th>Feature</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\langle +\text{subj} -t, -\text{subj} \rangle_0$ → $\langle +\text{subj} -t \rangle_0 \langle -\text{subj} \rangle_0$</td>
<td>$\exp(\lambda_{\text{INSERT}})$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\langle +\text{subj} -t, -\text{subj} \rangle_0$ → $\langle +\text{subj} -t \rangle_0 \langle -\text{subj} \rangle_1$</td>
<td>$\exp(\lambda_{\text{INSERT}})$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\langle +\text{subj} -t, -\text{subj} \rangle_0$ → $\langle +v +\text{subj} -t, -v, -\text{subj} \rangle_1$</td>
<td>$\exp(\lambda_{\text{MRG}} + \lambda_v)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Wh-phrase same as subject or separated because of doubling?

<table>
<thead>
<tr>
<th>Context</th>
<th>Transition</th>
<th>Feature</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\langle -t, -wh \rangle_0$ → $\langle +\text{subj} -t, -\text{subj} -\text{wh} \rangle_0$</td>
<td>$\exp(\lambda_{\text{MRG}} + \lambda_{\text{subj}})$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\langle -t, -wh \rangle_0$ → $\langle +\text{subj} -t, -\text{subj}, -\text{wh} \rangle_0$</td>
<td>$\exp(\lambda_{\text{MRG}} + \lambda_{\text{subj}})$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Learned weights on the MG

\[
\begin{align*}
\lambda_t &= 0.094350 \quad \exp(\lambda_t) = 1.0989 \\
\lambda_{\text{subj}} &= -5.734063 \quad \exp(\lambda_{\text{subj}}) = 0.0032 \\
\lambda_{\text{wh}} &= -0.094350 \quad \exp(\lambda_{\text{wh}}) = 0.9100 \\
\lambda_{\text{MERGE}} &= 0.629109 \quad \exp(\lambda_{\text{MERGE}}) = 1.8759 \\
\lambda_{\text{MOVE}} &= -0.629109 \quad \exp(\lambda_{\text{MOVE}}) = 0.5331 \\
\end{align*}
\]

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

\[
P(\text{wh is reflexivized}) = \frac{\exp(\lambda_{\text{MOVE}})}{\exp(\lambda_{\text{MERGE}}) + \exp(\lambda_{\text{MOVE}})} = 0.2213
\]

\[
P(\text{wh not reflexivized}) = \frac{\exp(\lambda_{\text{MERGE}})}{\exp(\lambda_{\text{MERGE}}) + \exp(\lambda_{\text{MOVE}})} = 0.7787
\]

\[
P(\text{question}) = \frac{\exp(\lambda_{\text{MOVE}} + \lambda_{\text{wh}})}{\exp(\lambda_{\text{MERGE}} + \lambda_t) + \exp(\lambda_{\text{MOVE}} + \lambda_{\text{wh}})} = 0.1905
\]

\[
P(\text{non-question}) = \frac{\exp(\lambda_{\text{MERGE}} + \lambda_t)}{\exp(\lambda_{\text{MERGE}} + \lambda_t) + \exp(\lambda_{\text{MOVE}} + \lambda_{\text{wh}})} = 0.8095
\]

\[
P(\text{non-wh subject merged and complex}) = \frac{\exp(\lambda_{\text{MERGE}})}{\exp(\lambda_{\text{MERGE}}) + \exp(\lambda_{\text{MERGE}}) + \exp(\lambda_{\text{MOVE}})} = 0.4378
\]

\[
P(\text{non-wh subject merged and lexical}) = \frac{\exp(\lambda_{\text{MERGE}})}{\exp(\lambda_{\text{MERGE}}) + \exp(\lambda_{\text{MERGE}}) + \exp(\lambda_{\text{MOVE}})} = 0.4378
\]

\[
P(\text{non-wh subject moved}) = \frac{\exp(\lambda_{\text{MOVE}})}{\exp(\lambda_{\text{MERGE}}) + \exp(\lambda_{\text{MERGE}}) + \exp(\lambda_{\text{MOVE}})} = 0.1244
\]

\[
P(\text{who will shave}) = 0.1905 \times 0.7787 = 0.148
\]

\[
P(\text{boys will shave themselves}) = 0.5 \times 0.8095 \times 0.1244 = 0.050
\]
**Learned weights on the IMG**

- $\lambda_t = 0.723549 \quad \exp(\lambda_t) = 2.0617 \quad P(\text{antecedent is lexical}) = 0.5$
- $\lambda_v = 0.440585 \quad \exp(\lambda_v) = 1.5536 \quad P(\text{antecedent is non-lexical}) = 0.5$
- $\lambda_{wh} = -0.723459 \quad \exp(\lambda_{wh}) = 0.4850 \quad P(\text{wh-phrase reflexivized}) = 0.5$
- $\lambda_{\text{INSERT}} = 0.440585 \quad \exp(\lambda_{\text{INSERT}}) = 1.5536 \quad P(\text{wh-phrase non-reflexivized}) = 0.5$
- $\lambda_{\text{MRG}} = -0.440585 \quad \exp(\lambda_{\text{MRG}}) = 0.6437$

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

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

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

**Grammar:** MG, i.e. **MERGE** and **MOVE**

**Sentence:** ‘who will shave themselves’

<table>
<thead>
<tr>
<th>MG, i.e. <strong>MERGE</strong> and <strong>MOVE</strong></th>
<th>Surprisal</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.35478</td>
<td>boys will shave</td>
<td></td>
</tr>
<tr>
<td>0.35478</td>
<td>some boys will shave</td>
<td></td>
</tr>
<tr>
<td>0.14801</td>
<td>who will shave</td>
<td></td>
</tr>
<tr>
<td>0.05022</td>
<td>boys will shave themselves</td>
<td></td>
</tr>
<tr>
<td>0.05022</td>
<td>some boys will shave themselves</td>
<td></td>
</tr>
<tr>
<td>0.04199</td>
<td>who will shave themselves</td>
<td></td>
</tr>
</tbody>
</table>
Surprisal predictions

**Grammar:** MG, i.e. **MERGE** and **MOVE**

**Sentence:** ‘who will shave themselves’

<table>
<thead>
<tr>
<th></th>
<th>MG, i.e. <strong>MERGE</strong> and <strong>MOVE</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.35478</td>
<td>boys will shave</td>
</tr>
<tr>
<td>0.35478</td>
<td>some boys will shave</td>
</tr>
<tr>
<td>0.14801</td>
<td>who will shave</td>
</tr>
<tr>
<td>0.05022</td>
<td>boys will shave themselves</td>
</tr>
<tr>
<td>0.05022</td>
<td>some boys will shave themselves</td>
</tr>
<tr>
<td>0.04199</td>
<td>who will shave themselves</td>
</tr>
</tbody>
</table>

surprisal at ‘who’ = \(- \log P(W_1 = \text{who})\)
\[= - \log (0.15 + 0.04)\]
\[= - \log 0.19\]
\[= 2.4\]

surprisal at ‘themselves’ = \(- \log P(W_4 = \text{themselves} | W_1 = \text{who}, \ldots)\)
\[= - \log \frac{0.04}{0.15 + 0.04}\]
\[= - \log 0.21\]
\[= 2.2\]
Surprisal predictions

**Grammar:** IMG, i.e. **MRG** and **INSERT**

**Sentence:** ‘who will shave themselves’

<table>
<thead>
<tr>
<th>IMG, i.e. <strong>MRG</strong> and <strong>INSERT</strong></th>
<th>surprisal</th>
</tr>
</thead>
<tbody>
<tr>
<td>boys will shave</td>
<td>0.35721</td>
</tr>
<tr>
<td>some boys will shave</td>
<td>0.35721</td>
</tr>
<tr>
<td>who will shave</td>
<td>0.095</td>
</tr>
<tr>
<td>who will shave themselves</td>
<td>0.095</td>
</tr>
<tr>
<td>boys will shave themselves</td>
<td>0.04779</td>
</tr>
<tr>
<td>some boys will shave themselves</td>
<td>0.04779</td>
</tr>
</tbody>
</table>
Surprisal predictions

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

surprisal at ‘who’ = \(- \log P(W_1 = \text{who})\)
= \(- \log(0.10 + 0.10)\)
= \(- \log 0.2\)
= 2.3

surprisal at ‘themselves’ = \(- \log P(W_4 = \text{themselves} | W_1 = \text{who}, \ldots)\)
= \(- \log \frac{0.10}{0.10 + 0.10}\)
= \(- \log 0.5\)
= 1
What are grammars? Derivationally distinct implementations of merge

Telling them apart

Historical perspective

MOVE
MERGE
C
did
John ... eat ... what

CP
who
C
TP
T
will
shave
DP
themselves

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 some boys will shave

MG, i.e. MERGE and MOVE
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.35721</td>
<td>boys will shave</td>
</tr>
<tr>
<td>0.35721</td>
<td>some boys will shave</td>
</tr>
<tr>
<td>0.14801</td>
<td>who will shave</td>
</tr>
<tr>
<td>0.095</td>
<td>who will shave themselves</td>
</tr>
<tr>
<td>0.04779</td>
<td>boys will shave themselves</td>
</tr>
<tr>
<td>0.04779</td>
<td>some boys will shave themselves</td>
</tr>
</tbody>
</table>

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

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

3
2
1
0
who
c
will
did
shave
themselves

3
2
1
0
who
c
will
did
shave
themselves

33 / 40
What are grammars? Derivationally distinct implementations of merge Telling them apart Historical perspective

$P(D|\hat{G}_{\text{DET}}) = 2.44 \times 10^{-13}$

$P(D|\hat{G}_{\text{WH}}) = 4.94 \times 10^{-14}$

$\frac{P(D|\hat{G}_{\text{DET}})}{P(D|\hat{G}_{\text{WH}})} = 4.94$

$10$ boys will shave

$2$ boys will shave themselves

$3$ who will shave

$1$ who will shave themselves

$5$ foo boys will shave

$P(D|\hat{G}_{\text{DET}}) = 1.46 \times 10^{-13}$

$P(D|\hat{G}_{\text{WH}}) = 1.62 \times 10^{-13}$

$\frac{P(D|\hat{G}_{\text{DET}})}{P(D|\hat{G}_{\text{WH}})} = 0.901$
Outline

1. What are grammars?
2. Derivationally distinct implementations of merge
3. Telling them apart
4. Historical perspective
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)

(1) $T_E - T_R - T_P - T_{AD}$

(2) $T_E - T_D - T_{lo}$

(3) $- T_P$
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).
What are grammars? Derivationally distinct implementations of merge Telling them apart

Historical perspective

Full derivational histories

\[ S \]

\[ NP \]

\[ John \]

\[ VP \]

\[ is \]

\[ Adj \]

\[ \{easy/eager\} \]

\[ to \]

\[ VP \]

\[ please \]

\[ T_5: \] front embedded object, replacing ‘it’

\[ T_4: \] delete ‘for someone’

\[ T_3: \] delete object

\[ T_2: \] delete duplicate NP

\[ T_1: \] replace COMP
What are grammars?

Derivationally distinct implementations of merge:

Telling them apart:

Historical perspective:
(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 was a thief] did she deny?

(2) Which claim [that Mary made] did she deny?
(1)  *Which claim [that Mary\textsubscript{i} was a thief] did she\textsubscript{i} deny?*

(2)  Which claim [that Mary\textsubscript{i} made] did she\textsubscript{i} deny?

```
move wh-phrase
```

```
move wh-phrase
```

```
move wh-phrase
```

```
move wh-phrase
```

```
move wh-phrase
```

```
adjoin relative clause
```

```
did
```

```
deny which claim that Mary\textsubscript{i} was a thief
```

```
did
```

```
deny which claim that Mary\textsubscript{i} made
```

```
did
```

```
deny which claim
```

```
did
```

```
deny which claim
```

```
did
```

```
deny which claim that Mary\textsubscript{i} was a thief
```

```
did
```

```
deny which claim that Mary\textsubscript{i} made
```

```
did
```

```
deny which claim
```
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.


