Sharpening the empirical claims of generative syntax through formalization

Tim Hunter

University of Minnesota, Twin Cities

ESSLLI, August 2015

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

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

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

Sharpening the empirical claims of generative syntax
through formalization

Tim Hunter — ESSLLI, August 2015

Part 1

Grammars and Cognitive Hypotheses

Outline

Outline

Claims made by grammars

What are grammars used for?

- “Mostly” for accounting for acceptability judgements
- But there are other ways a grammar can figure in claims about cognition

Claims made by grammars

What are grammars used for?

- “Mostly” for accounting for acceptability judgements
- But there are other ways a grammar can figure in claims about cognition

Often tempting to draw a distinction between “linguistic evidence” (where grammar lives) and “experimental evidence” (where cognition lives)

- One need not make this distinction
- We will proceed without it, i.e. it’s all linguistic (and/or all experimental)

Claims made by grammars

There's a "boring" sense in which every syntax paper makes a cognitive claim, i.e. a claim testable via acceptability facts.

Claims made by grammars

There's a "boring" sense in which every syntax paper makes a cognitive claim, i.e. a claim testable via acceptability facts.

- For one thing, this is not a cop-out!
    - Why does it seem like a cop-out?
    - Lingering externalism/Platonism?
    - Perhaps partly because it's just relatively rare to see anything being tested by other measures

Claims made by grammars

There's a "boring" sense in which every syntax paper makes a cognitive claim, i.e. a claim testable via acceptability facts.

- For one thing, this is not a cop-out!
  - Why does it seem like a cop-out?
  - Lingering externalism/Platonism?
  - Perhaps partly because it's just relatively rare to see anything being tested by other measures
- For another, we can incorporate grammars into claims that are testable by other measures.
  - This is the main point of the course!
  - The claims/predictions will depend on internal properties of grammars, not just what they say is good and what they say is bad
  - And we'll do it without seeing grammatical derivations as real-time operations

Claims made by grammars

If we accept — as I do — . . . that the rules of grammar enter into the processing mechanisms, then evidence concerning production, recognition, recall, and language use in general can be expected (in principle) to have bearing on the investigation of rules of grammar, on what is sometimes called "grammatical competence" or "knowledge of language".

(Chomsky 1980: pp.200-201)

[S]ince a competence theory must be incorporated in a performance model, evidence about the actual organization of behavior may prove crucial to advancing the theory of underlying competence.

(Chomsky 1980: p.226)

Claims made by grammars

> *If we accept — as I do — . . . that the rules of grammar enter into the processing mechanisms, then evidence concerning production, recognition, recall, and language use in general can be expected (in principle) to have bearing on the investigation of rules of grammar, on what is sometimes called "grammatical competence" or "knowledge of language".*
>
> *(Chomsky 1980: pp.200-201)*

> *[S]ince a competence theory must be incorporated in a performance model, evidence about the actual organization of behavior may prove crucial to advancing the theory of underlying competence.*
>
> *(Chomsky 1980: p.226)*

Evidence about X can only advance Y if Y makes claims about X!

Preview

What we will do:

- Put together a chain of linking hypotheses that bring “experimental evidence” to bear on “grammar questions”
    - e.g. reading times, acquisition patterns
    - e.g. move as distinct operation from merge vs. unified with merge
- Illustrate with some toy examples

Preview

What we will do:

- Put together a chain of linking hypotheses that bring “experimental evidence” to bear on “grammar questions”
    - e.g. reading times, acquisition patterns
    - e.g. move as distinct operation from merge vs. unified with merge
- Illustrate with some toy examples

What we will not do:

- Engage with state-of-the-art findings in the sentence processing literature
- End up with claims that one particular set of derivational operations is empirically better than another

Teasers

We'll take pairs of equivalent grammars that differ only in the move/re-merge dimension.

- They will make different predictions about sentence comprehension difficulty.
- They will make different predictions about what a learner will conclude from a common input corpus.

Teasers

We'll take pairs of equivalent grammars that differ only in the move/re-merge dimension.

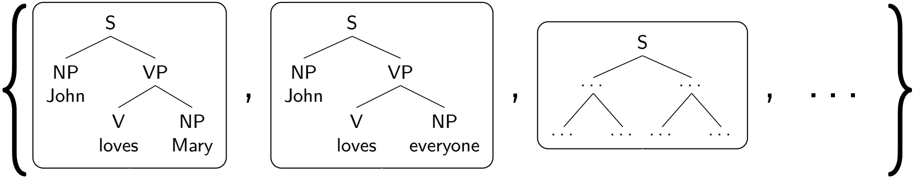
- They will make different predictions about sentence comprehension difficulty.
- They will make different predictions about what a learner will conclude from a common input corpus.

The issues become "distant but empirical questions". That's all we're aiming for, for now.

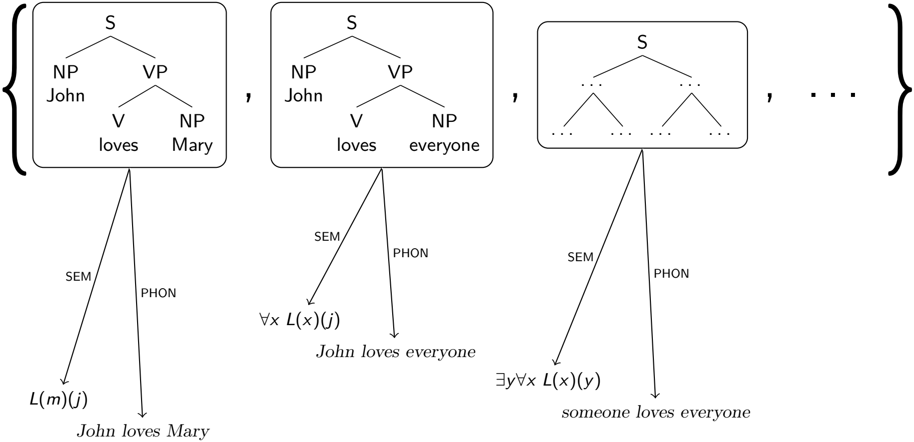
Outline

1. What we want to do with grammars
2. How to get grammars to do it
3. Derivations and representations
4. Information-theoretic complexity metrics

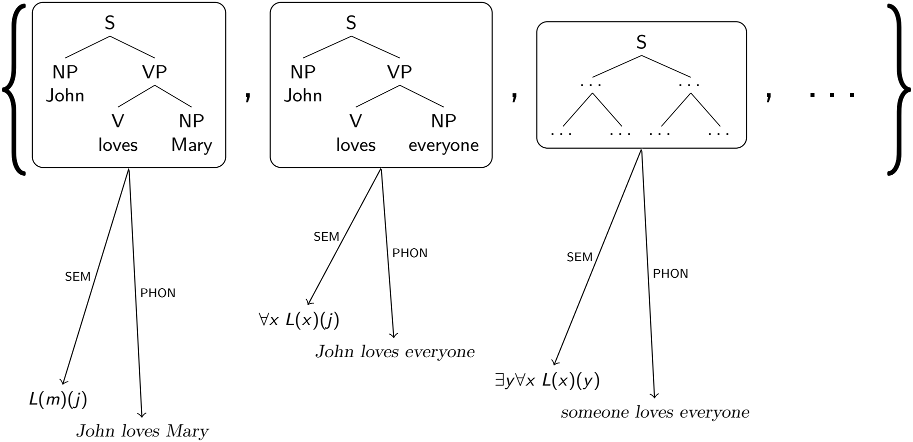
Interpretation functions



Interpretation functions



Interpretation functions



Caveats:

- Maybe we're interested in the finite specification of the set
- Maybe there's no clear line between observable and not
- Maybe some evidence is based on relativities among interpretations

Telling grammars apart

So, what if we have two different grammars — systems that define different sets of objects — that we can't tell apart via the sound and meaning interpretations?

(Perhaps because they're provably equivalent, or perhaps because the evidence just happens to be unavailable.)

Telling grammars apart

So, what if we have two different grammars — systems that define different sets of objects — that we can't tell apart via the sound and meaning interpretations?

(Perhaps because they're provably equivalent, or perhaps because the evidence just happens to be unavailable.)

- Option 1: Conclude that the differences are irrelevant to us (or "they're not actually different").
- Option 2: Make the differences matter . . . somehow . . .

What are syntactic representations for?

Morrill (1994) in favour of Option 1:

> *The construal of a language as a collection of signs [sound-meaning pairs] presents as an investigative task the characterisation of this collection. This is usually taken to mean the specification of a set of "structural descriptions" (or: "syntactic structures"). Observe however that on our understanding a sign is an association of prosodic [phonological] and semantic properties. It is these properties that can be observed and that are to be modelled. There appears to be no observation which bears directly on syntactic as opposed to prosodic and/or semantic properties, and this implies an asymmetry in the status of these levels.* ***A structural description is only significant insofar as it is understood as predicting prosodic and semantic properties (e.g. in interpreting the yield of a tree as word order). Attribution of syntactic (or prosodic or semantic) structure does not of itself predict anything.***

What are syntactic representations for?

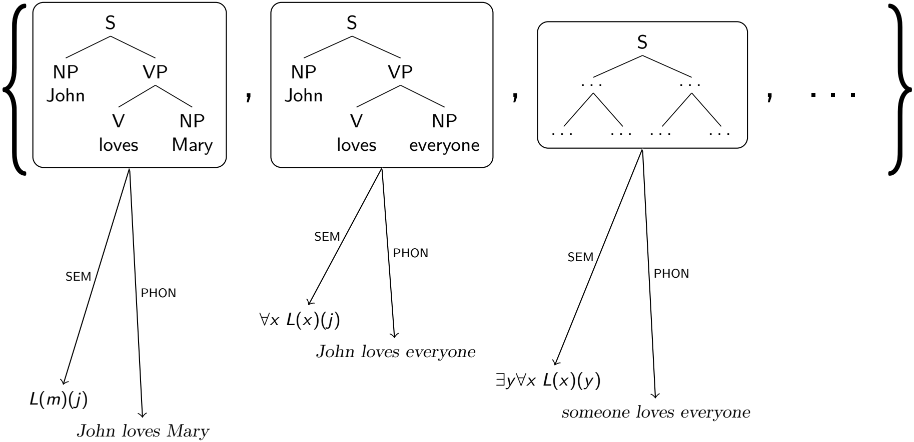
Morrill (1994) in favour of Option 1:

> *The construal of a language as a collection of signs [sound-meaning pairs] presents as an investigative task the characterisation of this collection. This is usually taken to mean the specification of a set of "structural descriptions" (or: "syntactic structures"). Observe however that on our understanding a sign is an association of prosodic [phonological] and semantic properties. It is these properties that can be observed and that are to be modelled. There appears to be no observation which bears directly on syntactic as opposed to prosodic and/or semantic properties, and this implies an asymmetry in the status of these levels.* ***A structural description is only significant insofar as it is understood as predicting prosodic and semantic properties (e.g. in interpreting the yield of a tree as word order). Attribution of syntactic (or prosodic or semantic) structure does not of itself predict anything.***

Where might we depart from this (to pursue Option 2)?

- Object that syntactic structure **does** matter "of itself"
- Object that prosodic and semantic properties are **not** the only ones we can observe

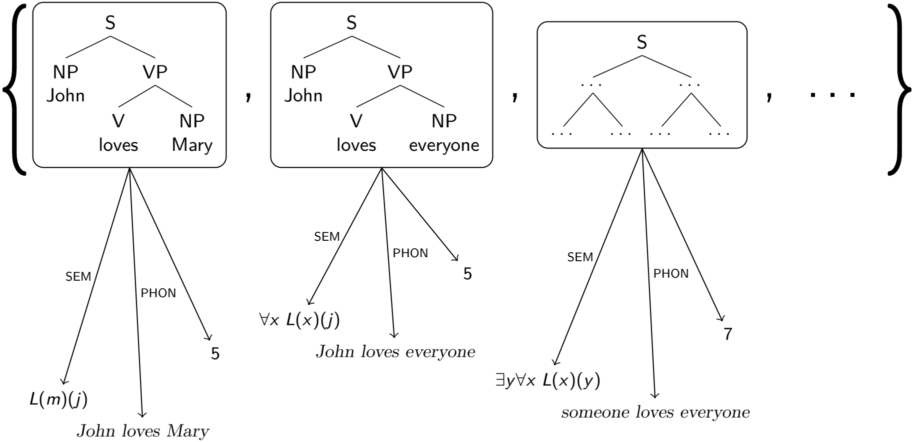
Interpretation functions



Caveats:

- Maybe we're interested in the finite specification of the set
- Maybe there's no clear line between observable and not
- Maybe some evidence is based on relativities among interpretations

Interpretation functions



Caveats:

- Maybe we're interested in the finite specification of the set
- Maybe there's no clear line between observable and not
- Maybe some evidence is based on relativities among interpretations

Interpretation functions for “complexity”

What are some other interpretation functions?

- number of nodes

Interpretation functions for “complexity”

What are some other interpretation functions?

- number of nodes
- ratio of total nodes to terminal nodes (Miller and Chomsky 1963)

Ratio of total nodes to terminal nodes

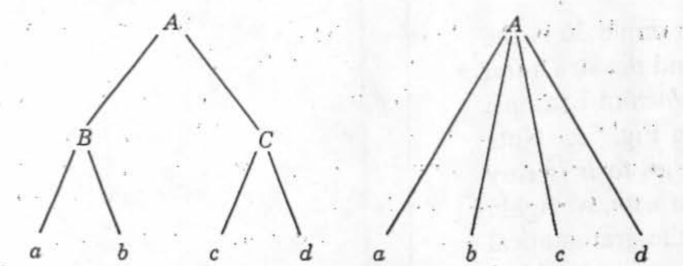


Fig. 8. Illustrating a measure of structural complexity. $N(Q)$ for the P-marker (a) is 7/4; for (b), $N(Q) = 5/4$.

(Miller and Chomsky 1963)

Ratio of total nodes to terminal nodes

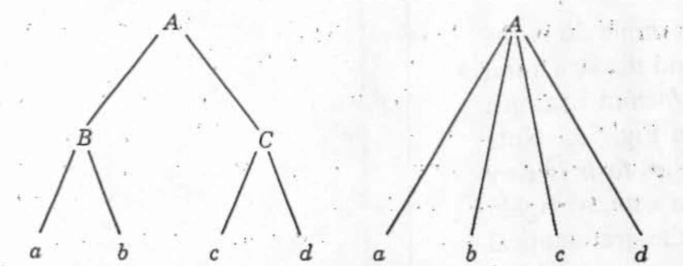


Fig. 8. Illustrating a measure of structural complexity. $N(Q)$ for the P-marker (a) is 7/4; for (b), $N(Q) = 5/4$.

Won't distinguish center-embedding from left- and right-embedding

(1) The mouse [the cat [the dog bit] chased] died. (center)

(2) The dog bit the cat [which chased the mouse [which died]]. (right)

(3) [[the dog] 's owner] 's friend (left)

(Miller and Chomsky 1963)

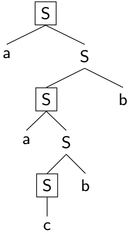
Interpretation functions for “complexity”

What are some other interpretation functions?

- number of nodes
- ratio of total nodes to terminal nodes (Miller and Chomsky 1963)
- degree of self-embedding (Miller and Chomsky 1963)

Degree of (centre-)self-embedding

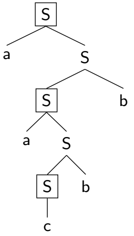
A tree's degree of self-embedding is m iff:
"there is . . . a continuous path passing through $m + 1$ nodes $N_0, \ldots, N_m$, each with the same label, where each N_i $(i \geq 1)$ is fully self-embedded (with something to the left and something to the right) in the subtree dominated by N_{i-1}"

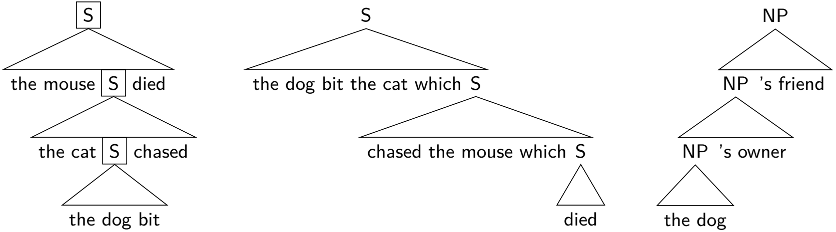


(Miller and Chomsky 1963)

Degree of (centre-)self-embedding

A tree's degree of self-embedding is m iff:
"there is . . . a continuous path passing through $m+1$ nodes $N_0, \ldots, N_m$, each with the same label, where each N_i ($i \geq 1$) is fully self-embedded (with something to the left and something to the right) in the subtree dominated by N_{i-1}"





(Miller and Chomsky 1963)

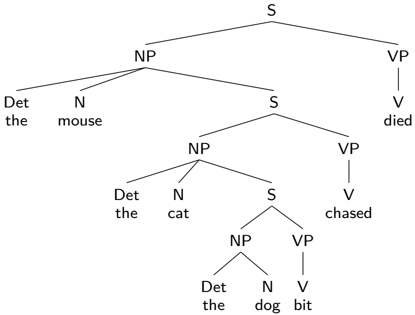
Interpretation functions for “complexity”

What are some other interpretation functions?

- number of nodes
- ratio of total nodes to terminal nodes (Miller and Chomsky 1963)
- degree of self-embedding (Miller and Chomsky 1963)
- “depth” of memory required by a top-down parser (Yngve 1960)

Yngve's depth

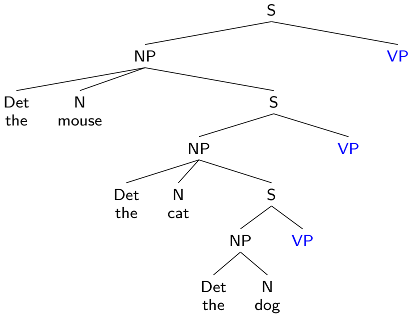
Number of constituents expected but not yet started:



(Yngve 1960)

Yngve's depth

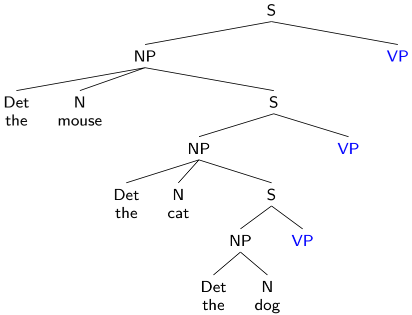
Number of constituents expected but not yet started:



(Yngve 1960)

Yngve's depth

Number of constituents expected but not yet started:

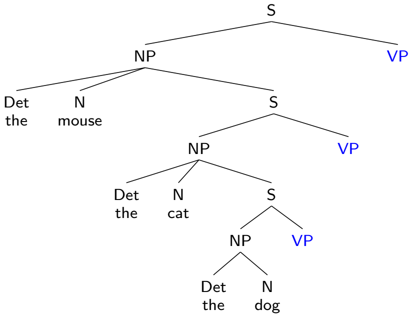


- Unlike (center-)self-embedding, right-embedding doesn't create such large lists of expected constituents (because the expected stuff is all part of **one** constituent).
- But left-embedding does.

(Yngve 1960)

Yngve's depth

Number of constituents expected but not yet started:



- Unlike (center-)self-embedding, right-embedding doesn't create such large lists of expected constituents (because the expected stuff is all part of **one** constituent).
- But left-embedding does.

- Yngve's theory was set within — perhaps justified by — a procedural story, but we can arguably detach it from that and treat depth as just another property of trees.

(Yngve 1960)

Interpretation functions for “complexity”

What are some other interpretation functions?

- number of nodes
- ratio of total nodes to terminal nodes (Miller and Chomsky 1963)
- degree of self-embedding (Miller and Chomsky 1963)
- “depth” of memory required by a top-down parser (Yngve 1960)
- minimal attachment, late closure, etc.? (Frazier and Clifton 1996)

Reaching conclusions about grammars

complexity metric + grammar ⟶ prediction

Typically, arguments hold the grammar fixed and present evidence in favour of a metric.

Reaching conclusions about grammars

complexity metric + grammar ⟶ prediction

Typically, arguments hold the grammar fixed and present evidence in favour of a metric.

We can flip this around: hold the metric fixed and present evidence in favour of a grammar.

> *If we accept — as I do — . . . that the rules of grammar enter into the processing mechanisms, then evidence concerning production, recognition, recall, and language use in general can be expected (in principle) to have bearing on the investigation of rules of grammar, on what is sometimes called "grammatical competence" or "knowledge of language".*
>
> *(Chomsky 1980: pp.200-201)*

Reaching conclusions about grammars

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

Example: hold self-embedding fixed as the complexity metric.

Reaching conclusions about grammars

complexity metric $+$ grammar $\longrightarrow$ prediction

Example: hold self-embedding fixed as the complexity metric.

(4) That [the food that [John ordered] tasted good] pleased him.

(5) That [that [the food was good] pleased John] surprised Mary.

Grammar question: Does a relative clause have a node labeled S?

Reaching conclusions about grammars

complexity metric + grammar ⟶ prediction

Example: hold self-embedding fixed as the complexity metric.

(4) That [the food that [John ordered] tasted good] pleased him.

(5) That [that [the food was good] pleased John] surprised Mary.

Grammar question: Does a relative clause have a node labeled S?

Proposed answer	(4) structure	(5) structure	Prediction
Yes	$\ldots[_S \ldots[_S \ldots]\ ]$	$\ldots[_S \ldots[_S \ldots]\ ]$	(4) & (5) same
No	$\ldots[_S \ldots[_{RC} \ldots]\ ]$	$\ldots[_S \ldots[_S \ldots]\ ]$	(5) harder

Reaching conclusions about grammars

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

Example: hold self-embedding fixed as the complexity metric.

(4) That [the food that [John ordered] tasted good] pleased him.

(5) That [that [the food was good] pleased John] surprised Mary.

Grammar question: Does a relative clause have a node labeled S?

Proposed answer	(4) structure	(5) structure	Prediction
Yes	$\ldots[_S \ldots [_S \ldots] \,]$	$\ldots[_S \ldots [_S \ldots] \,]$	(4) & (5) same
No	$\ldots[_S \ldots [_{RC} \ldots] \,]$	$\ldots[_S \ldots [_S \ldots] \,]$	(5) harder

Conclusion: The fact that (5) is harder supports the “No” answer.

Outline

1. What we want to do with grammars

2. How to get grammars to do it

3. Derivations and representations

4. Information-theoretic complexity metrics

Derivations and representations

Question

But these metrics are all properties of a final, fully-constructed tree.
How can anything like this be sensitive to differences in the derivational operations that build these trees? (e.g. TAG vs. MG, whether move is re-merge)

Interpretation functions for "complexity"

What are some other interpretation functions?

- number of nodes
- ratio of total nodes to terminal nodes (Miller and Chomsky 1963)
- degree of self-embedding (Miller and Chomsky 1963)
- "depth" of memory required by a top-down parser (Yngve 1960)
- minimal attachment, late closure, etc.? (Frazier and Clifton 1996)
- "nature, number and complexity of" transformations (Miller and Chomsky 1963)

“nature, number and complexity of the grammatical transformations involved”

> *The psychological plausibility of a transformational model of the language user would be strengthened, of course, if it could be shown that our performance on tasks requiring an appreciation of the structure of transformed sentences is* ***some function of the nature, number and complexity of the grammatical transformations involved****.*
>
> *(Miller and Chomsky 1963: p.481)*

Derivations and representations

Question

But these metrics are all properties of a final, fully-constructed tree.
How can anything like this be sensitive to differences in the derivational operations that build these trees? (e.g. TAG vs. MG, whether move is re-merge)

Derivations and representations

Question

But these metrics are all properties of a final, fully-constructed tree.
How can anything like this be sensitive to differences in the derivational operations that build these trees? (e.g. TAG vs. MG, whether move is re-merge)

Answer

The relevant objects on which the interpretation functions are defined encode a complete derivational history.

Derivations and representations

Question

But these metrics are all properties of a final, fully-constructed tree.
How can anything like this be sensitive to differences in the derivational operations that build these trees? (e.g. TAG vs. MG, whether move is re-merge)

Answer

The relevant objects on which the interpretation functions are defined encode a complete derivational history.

e.g. The function which, given a complete "recipe" for carrying out a derivation, returns the number of movement steps called for by the recipe.

Full derivation recipes?

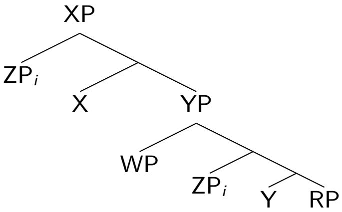
Are the inputs to these functions **really** full derivation recipes?

For minimalist syntax it's hard to tell, because the final derived object very often uniquely identifies a derivational history/recipe.

Full derivation recipes?

Are the inputs to these functions **really** full derivation recipes?

For minimalist syntax it's hard to tell, because the final derived object very often uniquely identifies a derivational history/recipe.



- merge Y with RP
- merge the result with ZP
- merge the result with WP
- merge X with the result
- move ZP

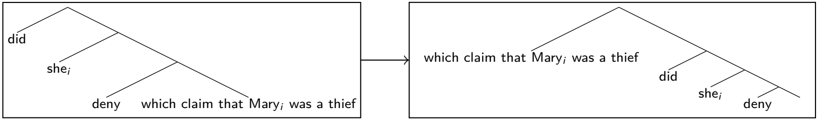
Full derivation recipes?

A few cases reveal that (we must all be already assuming that) it's full derivations/recipes that count.

Full derivation recipes?

A few cases reveal that (we must all be already assuming that) it's full derivations/recipes that count.

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

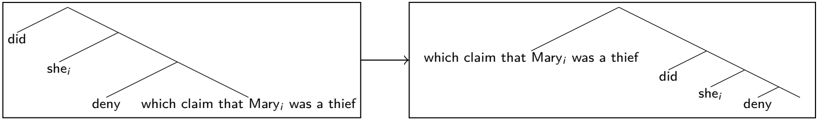


(7) Which claim [that Mary$_i$ made] did she$_i$ deny?

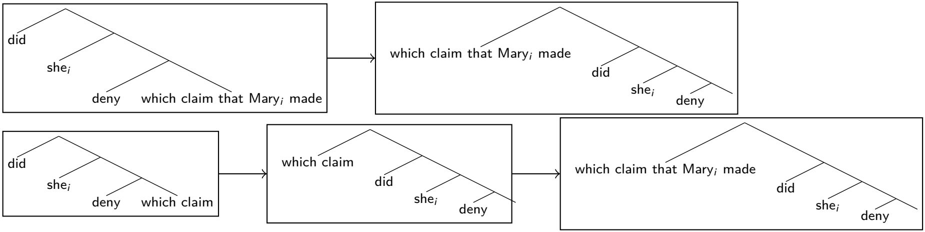
Full derivation recipes?

A few cases reveal that (we must all be already assuming that) it's full derivations/recipes that count.

(6) *Which claim [that Mary_i was a thief] did she_i deny?



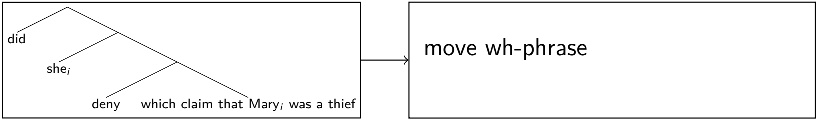
(7) Which claim [that Mary_i made] did she_i deny?



Full derivation recipes?

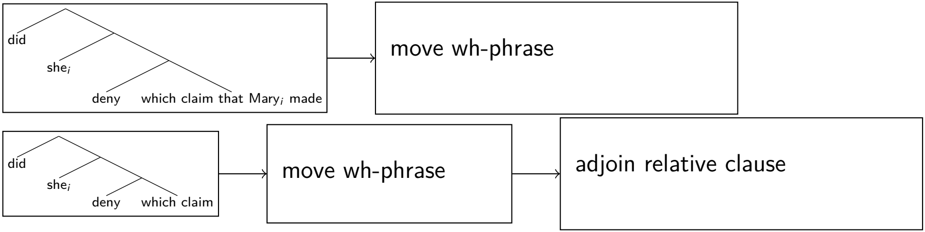
A few cases reveal that (we must all be already assuming that) it's full derivations/recipes that count.

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



move wh-phrase

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



move wh-phrase

move wh-phrase

adjoin relative clause

Full derivation recipes?

Also:

- subjacency effects without traces
- compare categorial grammar

Full derivation recipes?

Also:

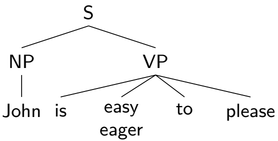
- subjacency effects without traces
- compare categorial grammar

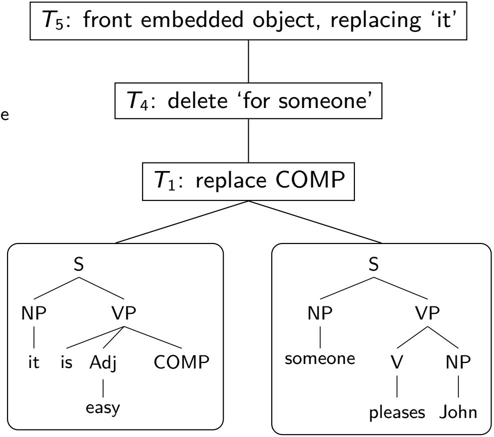
And this is not a new idea!

> *[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)*

Full derivation recipes?

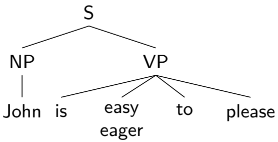
> *[The perceptual model] will utilize the full resources of the transformational grammar to provide a structural description, consisting of a set of P-markers and a transformational history*
>
> *Miller and Chomsky (1963: p.480)*

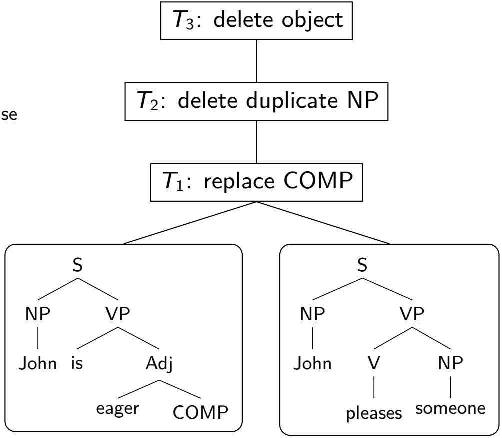




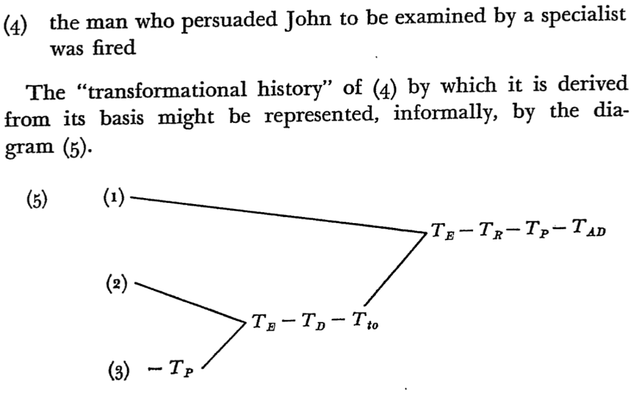
Full derivation recipes?

> *[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)*





Full derivation recipes?



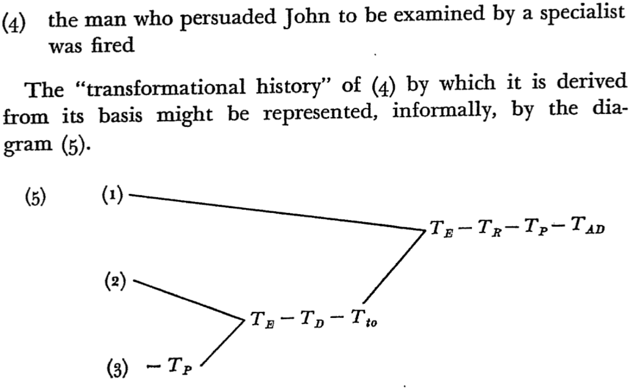
(4) the man who persuaded John to be examined by a specialist was fired

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

(5)

(Chomsky 1965)

Full derivation recipes?



(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)

Differences these days:

- We'll have things like merge and move at the internal nodes instead of T_P, T_E, etc.
- We'll have lexical items at the leaves rather than base-derived trees.

(Chomsky 1965)

Outline

1. What we want to do with grammars
2. How to get grammars to do it
3. Derivations and representations
4. Information-theoretic complexity metrics

Surprisal and entropy reduction

Why these complexity metrics?

- Partly just for concreteness, to give us a goal.
- They are formalism neutral to a degree that others aren't.
- They are mechanism neutral (Marr level one).
- The pieces of the puzzle that we need to get there (e.g. probabilities) seem likely to be usable in other ways.

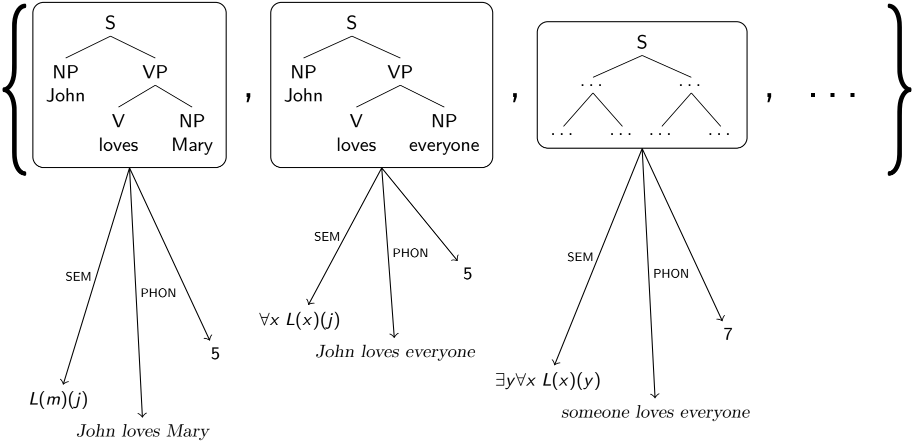
Surprisal and entropy reduction

Why these complexity metrics?

- Partly just for concreteness, to give us a goal.
- They are formalism neutral to a degree that others aren't.
- They are mechanism neutral (Marr level one).
- The pieces of the puzzle that we need to get there (e.g. probabilities) seem likely to be usable in other ways.

John Hale, Cornell Univ.

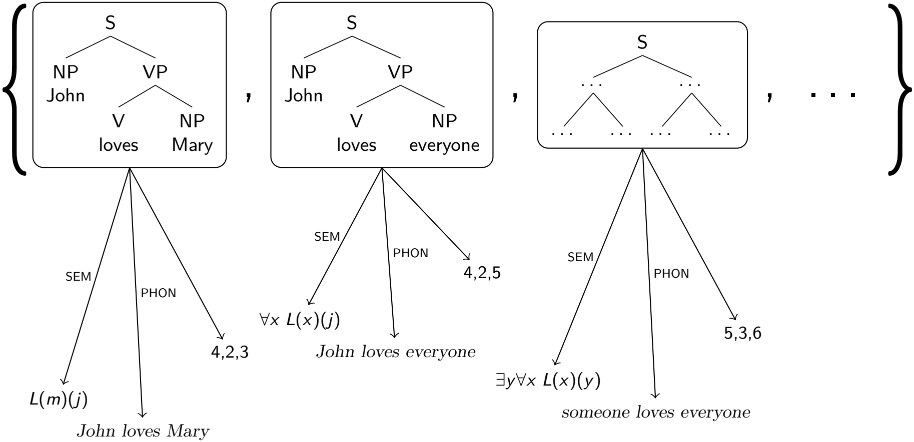
Interpretation functions



Caveats:

- Maybe we're interested in the finite specification of the set
- Maybe there's no clear line between observable and not
- Maybe some evidence is based on relativities among interpretations

Interpretation functions



Caveats:

- Maybe we're interested in the finite specification of the set
- Maybe there's no clear line between observable and not
- Maybe some evidence is based on relativities among interpretations

Surprisal

Given a sentence $w_1 w_2 \ldots w_n$:

$$\text{surprisal at } w_i = -\log P(W_i = w_i \mid W_1 = w_1, W_2 = w_2, \ldots, W_{i-1} = w_{i-1})$$

Surprisal

0.4	John ran
0.15	John saw it
0.05	John saw them
0.25	Mary ran
0.1	Mary saw it
0.05	Mary saw them

What predictions can we make about the difficulty of comprehending
'John saw it'?

Surprisal

0.4	John ran
0.15	John saw it
0.05	John saw them
0.25	Mary ran
0.1	Mary saw it
0.05	Mary saw them

What predictions can we make about the difficulty of comprehending 'John saw it'?

$$\begin{aligned}\text{surprisal at 'John'} &= -\log P(W_1 = \text{John}) \\ &= -\log(0.4 + 0.15 + 0.05) \\ &= -\log 0.6 \\ &= 0.74\end{aligned}$$

Surprisal

0.4	John ran
0.15	John saw it
0.05	John saw them
0.25	Mary ran
0.1	Mary saw it
0.05	Mary saw them

What predictions can we make about the difficulty of comprehending 'John saw it'?

$$\begin{aligned}\text{surprisal at 'John'} &= -\log P(W_1 = \text{John}) \\ &= -\log(0.4 + 0.15 + 0.05) \\ &= -\log 0.6 \\ &= 0.74\end{aligned}$$

$$\begin{aligned}\text{surprisal at 'saw'} &= -\log P(W_2 = \text{saw} \mid W_1 = \text{John}) \\ &= -\log \frac{0.15 + 0.05}{0.4 + 0.15 + 0.05} \\ &= -\log 0.33 \\ &= 1.58\end{aligned}$$

Surprisal

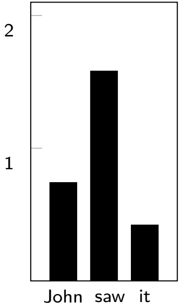
0.4	John ran
0.15	John saw it
0.05	John saw them
0.25	Mary ran
0.1	Mary saw it
0.05	Mary saw them

What predictions can we make about the difficulty of comprehending 'John saw it'?

$$\begin{aligned}\text{surprisal at 'John'} &= -\log P(W_1 = \text{John})\\ &= -\log(0.4 + 0.15 + 0.05)\\ &= -\log 0.6\\ &= 0.74\end{aligned}$$

$$\begin{aligned}\text{surprisal at 'saw'} &= -\log P(W_2 = \text{saw} \mid W_1 = \text{John})\\ &= -\log \frac{0.15 + 0.05}{0.4 + 0.15 + 0.05}\\ &= -\log 0.33\\ &= 1.58\end{aligned}$$

$$\begin{aligned}\text{surprisal at 'it'} &= -\log P(W_3 = \text{it} \mid W_1 = \text{John}, W_2 = \text{saw})\\ &= -\log \frac{0.15}{0.15 + 0.05}\\ &= -\log 0.75\\ &= 0.42\end{aligned}$$



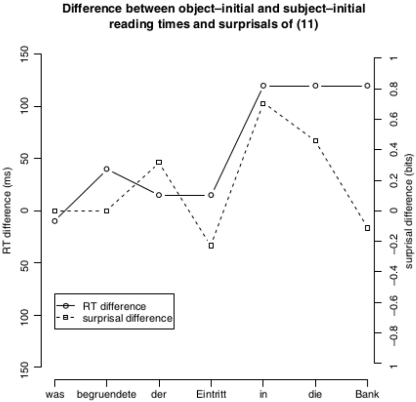
Accurate predictions made by surprisal

(Hale 2001)

Accurate predictions made by surprisal

(8) The reporter [who ____ attacked the senator] left the room. (easier)

(9) The reporter [who the senator attacked ____] left the room. (harder)


Difference between object–initial and subject–initial reading times and surprisals of (11)

(Levy 2008)

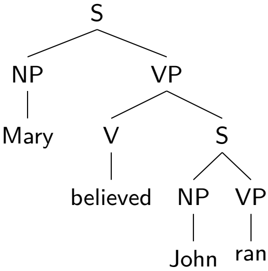
An important distinction

Using surprisal as a complexity metric says nothing about the form of the knowledge that the language comprehender is using!

- We're asking "what's the probability of w_i, given that we've seen $w_1 \dots w_{i-1}$ in the past".
- This does not mean that the comprehender's knowledge takes the form of answers to this kind of question.
- The linear nature of the metric reflects the task, not the knowledge being probed.

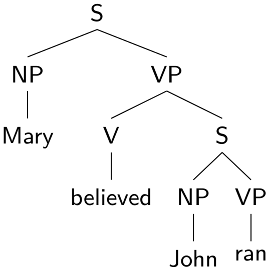
Probabilistic CFGs

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



Probabilistic CFGs

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



$$
\begin{aligned}
P(\text{Mary believed John ran}) &= 1.0 \times 0.7 \times 0.3 \times 0.4 \times 1.0 \times 0.3 \times 0.2 \\
&= 0.00504
\end{aligned}
$$

Surprisal with probabilistic CFGs

Goal: Calculate step-by-step surprisal values for ‘Mary believed John ran’

surprisal at ‘John’ $= -\log P(W_3 = \text{John} \mid W_1 = \text{Mary}, W_2 = \text{believed})$

Surprisal with probabilistic CFGs

Goal: Calculate step-by-step surprisal values for ‘Mary believed John ran’

surprisal at ‘John’ $= -\log P(W_3 = \text{John} \mid W_1 = \text{Mary}, W_2 = \text{believed})$

0.098	**Mary believed** Mary
0.042	**Mary believed** John
0.012348	**Mary believed** Mary knew Mary
0.01176	**Mary believed** Mary ran
0.008232	**Mary believed** Mary believed Mary
0.005292	**Mary believed** Mary knew John
0.005292	**Mary believed** John knew Mary
0.00504	**Mary believed** John ran
...	...

Surprisal with probabilistic CFGs

Goal: Calculate step-by-step surprisal values for ‘Mary believed John ran’

surprisal at ‘John’ $= -\log P(W_3 = \text{John} \mid W_1 = \text{Mary}, W_2 = \text{believed})$

0.098	**Mary believed** Mary
0.042	**Mary believed** John
0.012348	**Mary believed** Mary knew Mary
0.01176	**Mary believed** Mary ran
0.008232	**Mary believed** Mary believed Mary
0.005292	**Mary believed** Mary knew John
0.005292	**Mary believed** John knew Mary
0.00504	**Mary believed** John ran
. . .	. . .

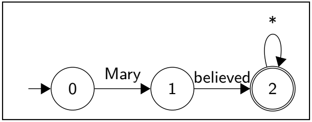
There are an infinite number of derivations consistent with input at each point!

$$\begin{aligned}\text{surprisal at ‘John’} &= -\log P(W_3 = \text{John} \mid W_1 = \text{Mary}, W_2 = \text{believed}) \\ &= -\log \frac{0.042 + 0.005292 + 0.00504 + \ldots}{0.098 + 0.042 + 0.12348 + 0.01176 + 0.008232 + \ldots}\end{aligned}$$

Intersection grammars

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

$\cap$

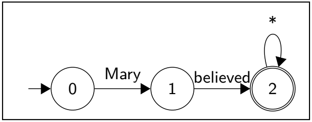


$= G_2$

Intersection grammars

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

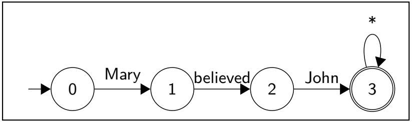
$\cap$



$= G_2$

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

$\cap$

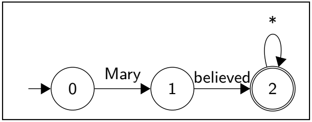


$= G_3$

Intersection grammars

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

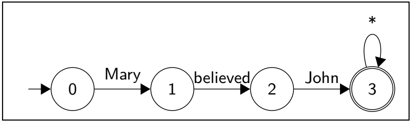
$\cap$



$= G_2$

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

$\cap$

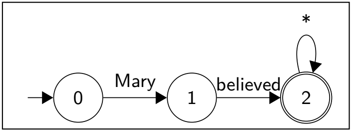


$= G_3$

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

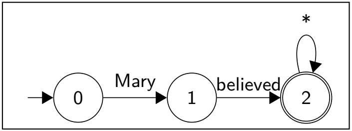
Grammar intersection example (simple)

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



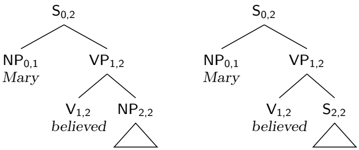
Grammar intersection example (simple)

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



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

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



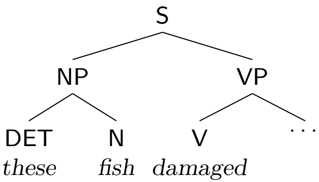
NB: Total weight in this grammar is not one! (What is it? Start symbol is $S_{0,2}$.)
Each derivation has the weight "it" had in the original grammar.

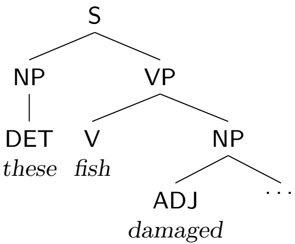
Grammar intersection example (more complicated)

S → NP VP
VP → V NP
NP → DET
NP → DET N
NP → ADJ N

V → *fish*
V → *damaged*
DET → *these*
N → *fish*
ADJ → *damaged*

These fish damaged . . .

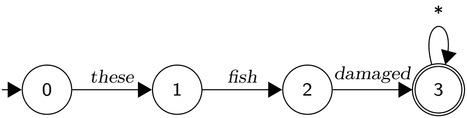




Grammar intersection example (more complicated)

S → NP VP
VP → V NP
NP → DET
NP → DET N
NP → ADJ N

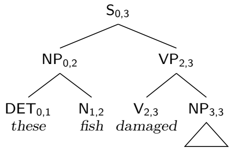
V → *fish*
V → *damaged*
DET → *these*
N → *fish*
ADJ → *damaged*

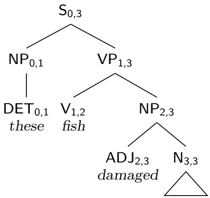


$\mathrm{S}_{0,3} \rightarrow \mathrm{NP}_{0,2}\ \mathrm{VP}_{2,3}$
$\mathrm{NP}_{0,2} \rightarrow \mathrm{DET}_{0,1}\ \mathrm{N}_{1,2}$
$\mathrm{VP}_{2,3} \rightarrow \mathrm{V}_{2,3}\ \mathrm{NP}_{3,3}$
$\mathrm{DET}_{0,1} \rightarrow \textit{these}$
$\mathrm{N}_{1,2} \rightarrow \textit{fish}$
$\mathrm{V}_{2,3} \rightarrow \textit{damaged}$

$\mathrm{S}_{0,3} \rightarrow \mathrm{NP}_{0,1}\ \mathrm{VP}_{1,3}$
$\mathrm{NP}_{0,1} \rightarrow \mathrm{DET}_{0,1}$
$\mathrm{VP}_{1,3} \rightarrow \mathrm{V}_{1,2}\ \mathrm{NP}_{2,3}$
$\mathrm{NP}_{2,3} \rightarrow \mathrm{ADJ}_{2,3}\ \mathrm{N}_{3,3}$
$\mathrm{V}_{1,2} \rightarrow \textit{fish}$
$\mathrm{ADJ}_{2,3} \rightarrow \textit{damaged}$

$\mathrm{NP}_{3,3} \rightarrow \mathrm{ADJ}_{3,3}\ \mathrm{N}_{3,3}$
$\mathrm{NP}_{3,3} \rightarrow \mathrm{DET}_{3,3}\ \mathrm{N}_{3,3}$
$\mathrm{NP}_{3,3} \rightarrow \mathrm{DET}_{3,3}$
$\mathrm{N}_{3,3} \rightarrow \textit{fish}$
$\mathrm{DET}_{3,3} \rightarrow \textit{these}$
$\mathrm{ADJ}_{3,3} \rightarrow \textit{damaged}$

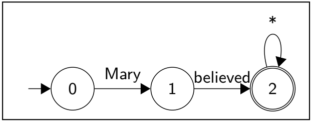




Intersection grammars

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

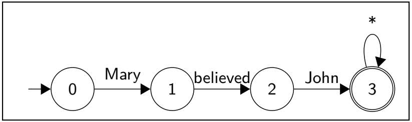
$\cap$



$= G_2$

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

$\cap$



$= G_3$

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

Computing sum of weights in a grammar (“partition function”)

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

(Nederhof and Satta 2008)

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

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

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

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

(Nederhof and Satta 2008)

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

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

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

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

Things to know

Technical facts about CFGs:

- Can intersect with a “prefix FSA”
- Can compute the total weight (and the entropy)

(Hale 2006)

Things to know

Technical facts about CFGs:

- Can intersect with a “prefix FSA”
- Can compute the total weight (and the entropy)

More generally:

- Intersecting a grammar with a prefix produces a new grammar which is a representation of the comprehender’s sentence-medial state
- So we can construct a sequence of grammars which represents the comprehender’s sequence of knowledge-states
- Ask “what changes” (or “how much changes”, etc.) at each step

The general approach is compatible with many very different grammar formalisms (any grammar formalism?) — provided the technical tricks can be pulled off.

(Hale 2006)

Looking ahead

Wouldn't it be nice if we could do all that for minimalist syntax?

The average syntax paper shows illustrative derivations, not a fragment.

What would we need?

- An explicit characterization of the set of possible derivations
- A way to "intersect" that with a prefix
- A way to define probability distributions over the possibilities

This will require certain idealizations. (But what's new?)

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

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

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

References I

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

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

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

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

Frazier, L. and Clifton, C. (1996). *Construal*. MIT Press, Cambridge, MA.

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

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

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

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

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

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

References II

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

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

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

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

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

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

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

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

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

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

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

References III

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

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

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