Sharpening the empirical claims of generative syntax through formalization

Tim Hunter

University of Minnesota, Twin Cities

NASSLLI, June 2014
Part 1: Grammars and cognitive hypotheses
  What is a grammar?
  What can grammars do?
  Concrete illustration of a target: Surprisal

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

Part 5: Learning and wrap-up
  Something slightly different: Learning model
  Recap and open questions
Sharpening the empirical claims of generative syntax through formalization

Tim Hunter — NASSLLI, June 2014

Part 5

Learning and wrap-up
Motivating question

Components of a learner:

- A formalism ("toolkit") defines a space of grammars for a learner to choose from
- An updating algorithm defines a way to search through such a space (in response to provided input)
Motivating question

Components of a learner:

- A formalism ("toolkit") defines a space of grammars for a learner to choose from.
- An updating algorithm defines a way to search through such a space (in response to provided input).

Given two formalisms, F1 and F2, can we construct a learner which

- reaches one end-state when used with F1, and
- reaches a different end-state when used with F2?
Motivating question

Components of a learner:

- A formalism ("toolkit") defines a space of grammars for a learner to choose from
- An updating algorithm defines a way to search through such a space (in response to provided input)

Given two formalisms, F1 and F2, can we construct a learner which

- reaches one end-state when used with F1, and
- reaches a different end-state when used with F2?

With everything else held fixed:

- same (strong) generative capacity
- same updating algorithm
- same training data
Outline

18 Grammatical formalisms and learning

19 Learning with a given grammar

20 Learning with a choice of grammars

21 Conclusion
Outline

18Grammatical formalisms and learning

19Learning with a given grammar

20Learning with a choice of grammars

21Conclusion
A "good sentence vs. bad sentence" learner will treat these two formalisms equivalently — it won’t "see" the internal differences in how they generate what they generate. (Gibson and Wexler 1994)
A “good sentence vs. bad sentence” learner will treat these two formalisms equivalently — it won’t “see” the internal differences in how they generate what they generate.

(Gibson and Wexler 1994)

Q: How can we provide traction between the learning algorithm and the internals of each $G$?
A “good sentence vs. bad sentence” learner will treat these two formalisms equivalently — it won’t “see” the internal differences in how they generate what they generate.

(Gibson and Wexler 1994)

Q: How can we provide traction between the learning algorithm and the internals of each $G$?
A: Probabilities
A “good sentence vs. bad sentence” learner will treat these two formalisms equivalently — it won’t “see” the internal differences in how they generate what they generate.

(Gibson and Wexler 1994)

Q: How can we provide traction between the learning algorithm and the internals of each $G$?
A: Probabilities
Learning scenario

Training corpus: some combination of occurrences of the following.

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

● The learner knows correct analyses of these sentences, with ‘foo’ as a determiner.
● The learner must decide what probabilities to attach to these known sentences.
Grammatical formalisms and learning

Learning with a given grammar

Learning with a choice of grammars

Conclusion

MGs

\[ G \]

Training corpus:

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

IMGs

\[ G \]
Training corpus:

10 boys will shave
2 boys will shave themselves
3 who will shave
1 who will shave themselves
5 foo boys will shave
Training corpus:

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

Grammar’s distribution:

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

Grammar's distribution:

- 0.35721 boys will shave
- 0.35721 foo boys will shave
- 0.095 who will shave
- 0.095 who will shave themselves
- 0.04779 boys will shave themselves
- 0.04779 foo boys will shave themselves
Training corpus:

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

<table>
<thead>
<tr>
<th></th>
<th>Entropy</th>
<th>Entropy Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>who</td>
<td>0.76</td>
<td>1.33</td>
</tr>
<tr>
<td>will</td>
<td>0.76</td>
<td>0.00</td>
</tr>
<tr>
<td>shave</td>
<td>0.76</td>
<td>0.00</td>
</tr>
<tr>
<td>themselves</td>
<td>0.00</td>
<td>0.76</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Entropy</th>
<th>Entropy Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>who</td>
<td>1.00</td>
<td>1.28</td>
</tr>
<tr>
<td>will</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>shave</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>themselves</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Outline

18 Grammatical formalisms and learning

19 Learning with a given grammar

20 Learning with a choice of grammars

21 Conclusion
Learning scenario

Training corpus: some combination of occurrences of the following:

<table>
<thead>
<tr>
<th>boys will shave</th>
<th>boys will shave themselves</th>
</tr>
</thead>
<tbody>
<tr>
<td>who will shave</td>
<td>who will shave themselves</td>
</tr>
<tr>
<td>foo boys will shave</td>
<td></td>
</tr>
</tbody>
</table>
Learning scenario

Training corpus: some combination of occurrences of the following.

<table>
<thead>
<tr>
<th>boys will shave</th>
<th>boys will shave themselves</th>
</tr>
</thead>
<tbody>
<tr>
<td>who will shave</td>
<td>who will shave themselves</td>
</tr>
<tr>
<td>foo boys will shave</td>
<td></td>
</tr>
</tbody>
</table>

- The learner knows correct analyses of wh-movement and reflexives.
Learning scenario

Training corpus: some combination of occurrences of the following.

<table>
<thead>
<tr>
<th>boys will shave</th>
<th>boys will shave themselves</th>
</tr>
</thead>
<tbody>
<tr>
<td>who will shave</td>
<td>who will shave themselves</td>
</tr>
<tr>
<td>foo boys will shave</td>
<td></td>
</tr>
</tbody>
</table>

- The learner **knows** correct analyses of wh-movement and reflexives.
- The learner **must decide** how to analyze ‘foo’: determiner or wh-phrase?
Training corpus:

<table>
<thead>
<tr>
<th></th>
<th>MG-DET</th>
<th>MG-WH</th>
<th>IMG-DET</th>
<th>IMG-WH</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>boys will shave</td>
<td>boys will shave themselves</td>
<td>who will shave</td>
<td>who will shave themselves</td>
</tr>
</tbody>
</table>
Grammatical formalisms and learning

MGs

- $G$
- $\hat{g}_{DET}$

MG-DET

- $G$
- $\hat{g}_{WH}$

MG-WH

- $G$

IMGs

- $G$

IMG-DET

- $G$

IMG-WH

Training corpus:

- 5 boys will shave
- 5 boys will shave themselves
- 5 who will shave
- 5 who will shave themselves
- 5 foo boys will shave

\[
\frac{P(D|\hat{g}_{DET})}{P(D|\hat{g}_{WH})} = \frac{3.36 \times 10^{-18}}{4.48 \times 10^{-20}} = 75.0
\]
### Training corpus:

- 5 boys will shave
- 5 boys will shave themselves
- 5 who will shave
- 5 who will shave themselves
- 5 foo boys will shave

<table>
<thead>
<tr>
<th>MG-DET</th>
<th>MG-WH</th>
<th>IMG-DET</th>
<th>IMG-WH</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{g}^{\text{DET}} )</td>
<td>( \hat{g}^{\text{WH}} )</td>
<td>( \hat{g}^{\text{DET}} )</td>
<td>( \hat{g}^{\text{WH}} )</td>
</tr>
</tbody>
</table>

\[
\frac{P(D|\hat{g}^{\text{DET}})}{P(D|\hat{g}^{\text{WH}})} = \frac{3.36 \times 10^{-18}}{4.48 \times 10^{-20}} = 75.0
\]

\[
\frac{P(D|\hat{g}^{\text{DET}})}{P(D|\hat{g}^{\text{WH}})} = \frac{3.36 \times 10^{-18}}{2.45 \times 10^{-19}} = 13.7
\]
Training corpus:

18 boys will shave
3 boys will shave themselves
1 who will shave
1 who will shave themselves
1 foo boys will shave

\[
\frac{P(D|\hat{g}_{\text{DET}})}{P(D|\hat{g}_{\text{WH}})} = \frac{5.82 \times 10^{-14}}{7.27 \times 10^{-11}} = 0.000801
\]

\[
\frac{P(D|\hat{g}_{\text{DET}})}{P(D|\hat{g}_{\text{WH}})} = \frac{7.64 \times 10^{-14}}{6.85 \times 10^{-10}} = 0.000112
\]
Grammatical formalisms and learning

Learning with a given grammar

Learning with a choice of grammars

Conclusion

Training corpus:

1 boys will shave
1 boys will shave themselves
8 who will shave
8 who will shave themselves
8 foo boys will shave

\[
\frac{P(D|\hat{g}_{\text{DET}})}{P(D|\hat{g}_{\text{WH}})} = \frac{1.21 \times 10^{-17}}{7.70 \times 10^{-19}} = 15.7
\]

\[
\frac{P(D|\hat{g}_{\text{DET}})}{P(D|\hat{g}_{\text{WH}})} = \frac{3.46 \times 10^{-17}}{1.19 \times 10^{-16}} = 0.291
\]
Training corpus:

- 8 boys will shave
- 1 boys will shave themselves
- 12 who will shave
- 1 who will shave themselves
- 4 foo boys will shave

\[
\frac{P(D|\hat{g}_{\text{DET}})}{P(D|\hat{g}_{\text{WH}})} = \frac{2.83 \times 10^{-15}}{4.36 \times 10^{-20}} = 64900 \\
\frac{P(D|\hat{g}_{\text{DET}})}{P(D|\hat{g}_{\text{WH}})} = \frac{1.31 \times 10^{-17}}{1.75 \times 10^{-17}} = 0.749
\]
Training corpus:

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

\[
\frac{P(D|\hat{g}_{\text{DET}})}{P(D|\hat{g}_{\text{WH}})} = \frac{2.44 \times 10^{-13}}{4.94 \times 10^{-14}} = 4.94
\]

\[
\frac{P(D|\hat{g}_{\text{DET}})}{P(D|\hat{g}_{\text{WH}})} = \frac{1.46 \times 10^{-13}}{1.62 \times 10^{-13}} = 0.901
\]
Details of one interesting case

**MG-WH**

Feature weight: ant=0.000000
Feature weight: obj=0.000000
Feature weight: subj=0.306077
Feature weight: t=-0.895880
Feature weight: v=0.000000
Feature weight: wh=0.895880
Feature weight: merge=-0.000000
Feature weight: move=-0.000000
{t29: 0.5, t13_t4: 0.5}
{t28: 0.5, t13_t5: 0.5}
{t0_t14: 0.077, t21_t7: 0.462, t22: 0.462}

t0 : (:: =t c)
t4 : (:: subj)
t5 : (:: subj -wh)
t7 : (:: wh)
t13 : (:: =subj t)
t14 : (:: t)
t21 : (:: =wh c)
t22 : (:: +wh c;: -wh)
t28 : (:: +subj t;: -subj;: -wh)
t29 : (:: +subj t;: -subj)

**IMG-WH**

Feature weight: ant=0.000000
Feature weight: obj=0.000000
Feature weight: subj=-0.860545
Feature weight: t=-0.434630
Feature weight: v=-3.324996
Feature weight: wh=2.050275
Feature weight: insert=-0.563888
Feature weight: merge=0.563888
{t00130005: 0.5, t0028: 0.5}
{t0021_t0007: 0.333, t00010016: 0.667}
{t00000014: 0.077, t0022: 0.923}
{t0013_t0004: 0.900, t00110026: 0.100}

t00000014 : (:: +t -c;: -t)
t00010016 : (:: +t +wh -c;: -t;: -wh)
t0004 : (:: -subj)
t0007 : (:: -wh)
t00110026 : (:: +v +subj -t;: -v;: -subj)
t0013 : (:: +subj -t)
t00130005 : (:: +subj -t;: -subj -wh)
t0021 : (:: +wh -c)
t0022 : (:: +wh -c;: -wh)
t0028 : (:: +subj -t;: -subj;: -wh)
Outline

18 Grammatical formalisms and learning

19 Learning with a given grammar

20 Learning with a choice of grammars

21 Conclusion
What we’ve done (I hope)

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)

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)
What we’ve done (I hope)

There are ways to have “purely derivational” properties of formalisms make a difference to predictions about sentence processing complexity and generalization in learning
What we’ve done (I hope)

There are ways to have “purely derivational” properties of formalisms make a
difference to predictions about sentence processing complexity and generalization in learning

- ...without saying anything about real-time mental operations
- ... (let alone saying that things like MERGE and MOVE happen in real time).
What we’ve done (I hope)

There are ways to have “purely derivational” properties of formalisms make a difference to predictions about sentence processing complexity and generalization in learning

- ...without saying anything about real-time mental operations
- ... (let alone saying that things like MERGE and MOVE happen in real time).
- Instead, the derivation tree is the object to be recovered/identified.
What we’ve done (I hope)

There are ways to have “purely derivational” properties of formalisms make a difference to predictions about sentence processing complexity and generalization in learning

- …without saying anything about real-time mental operations
- …(let alone saying that things like `MERGE` and `MOVE` happen in real time).
- Instead, the derivation tree is the object to be recovered/identified.

As mentioned above, the MP as a syntactic theory appears to be a step backwards for psycholinguistics (although perhaps not for syntacticians, of course). One of the fundamental problems is that the model derives a tree starting from all the lexical items and working up to the top-most node, which obviously is difficult to reconcile with left-to-right incremental parsing

*Ferreira (2005: p.369)*
What we’ve done (I hope)

There are ways to have “purely derivational” properties of formalisms make a difference to predictions about sentence processing complexity and generalization in learning

- ...without saying anything about real-time mental operations
- ... (let alone saying that things like MERGE and MOVE happen in real time).
- Instead, the derivation tree is the object to be recovered/identified.

As mentioned above, the MP as a syntactic theory appears to be a step backwards for psycholinguistics (although perhaps not for syntacticians, of course). One of the fundamental problems is that the model derives a tree starting from all the lexical items and working up to the top-most node, which obviously is difficult to reconcile with left-to-right incremental parsing

Ferreira (2005: p.369)

- What we’ve done of course leaves questions about real-time operations unanswered.
- But it’s not clear that there is a conflict that needs to be “reconciled”.
Open questions

How realistic is the assumption that there are a finite number of derivational states?

- MGs’ SMC vs. mainstream “minimality”
- Dependencies over arbitrary distances (e.g. Condition C, NPIs)
- ...?

Local vs. global normalization


