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?

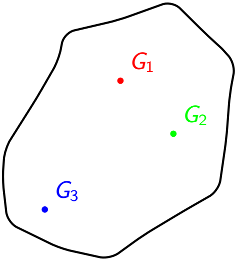
With everything else held fixed:

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

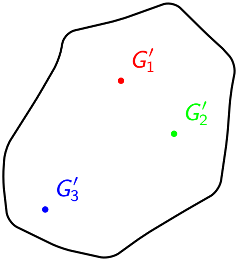
Outline

Outline

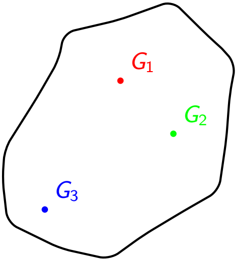
Formalism F1



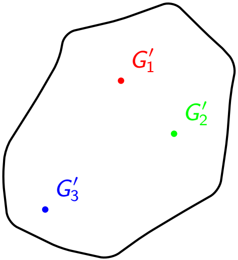
Formalism F2



Formalism F1



Formalism F2

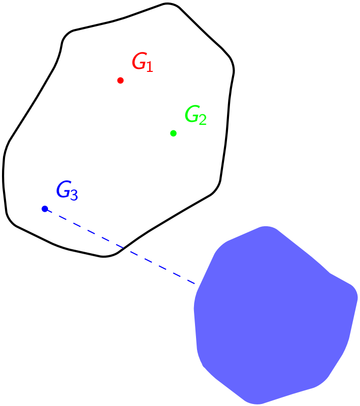


A "good sentence vs. bad sentence" learner will treat these two formalisms equivalently — it won't "see" the internal differences in how they generate what they generate.

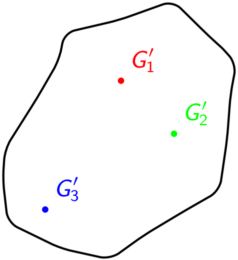
(Gibson and Wexler 1994)

Q: How can we provide traction between the learning algorithm and the internals of each G?

Formalism F1



Formalism F2

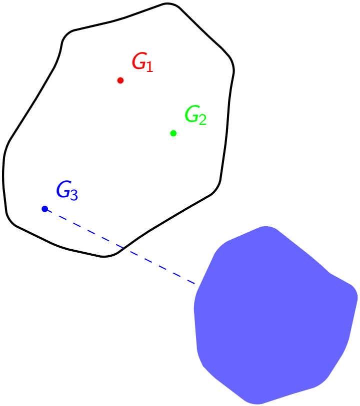


A "good sentence vs. bad sentence" learner will treat these two formalisms equivalently — it won't "see" the internal differences in how they generate what they generate.

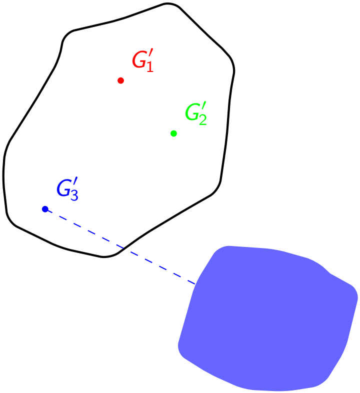
(Gibson and Wexler 1994)

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

Formalism F1



Formalism F2



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

Outline

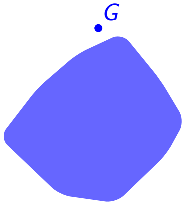
Learning scenario

Training corpus: some combination of occurrences of the following.

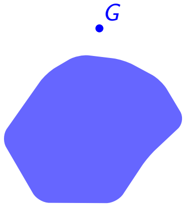
boys will shave	boys will shave themselves
who will shave	who will shave themselves
foo boys will shave	

- 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.

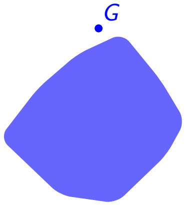
MGs



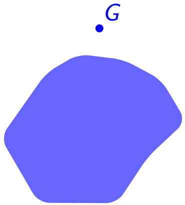
IMGs



MGs

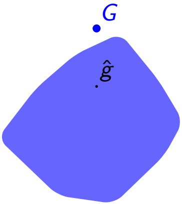
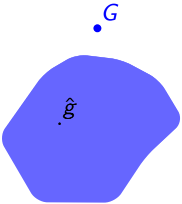


IMGs



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

MGs	IMGs
	

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

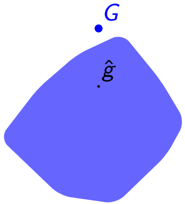
Grammar's distribution: (MGs)

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

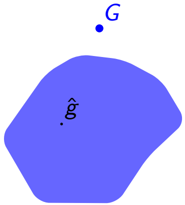
Grammar's distribution: (IMGs)

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

MGs



IMGs



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

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

	Entropy	Entropy Reduction
—	2.28	—
who	1.00	1.28
will	1.00	0.00
shave	1.00	0.00
themselves	0.00	1.00

Outline

Learning scenario

Training corpus: some combination of occurrences of the following.

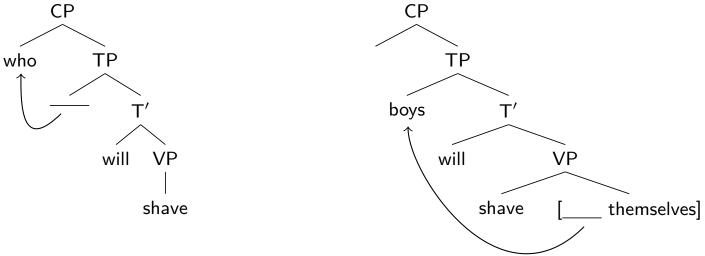
boys will shave	boys will shave themselves
who will shave	who will shave themselves
foo boys will shave	

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 wh-movement and reflexives.

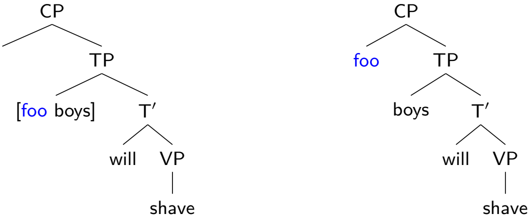


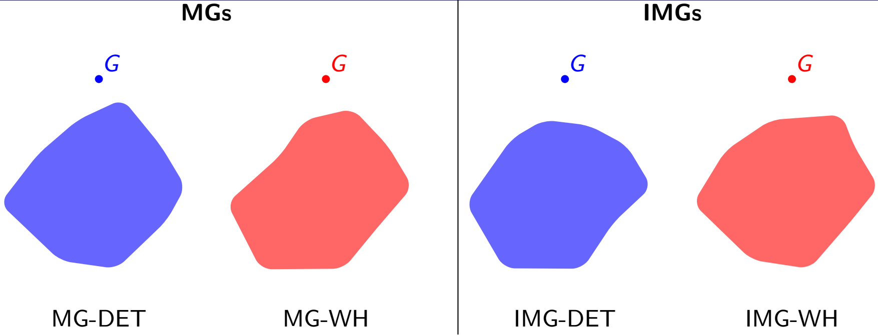
Learning scenario

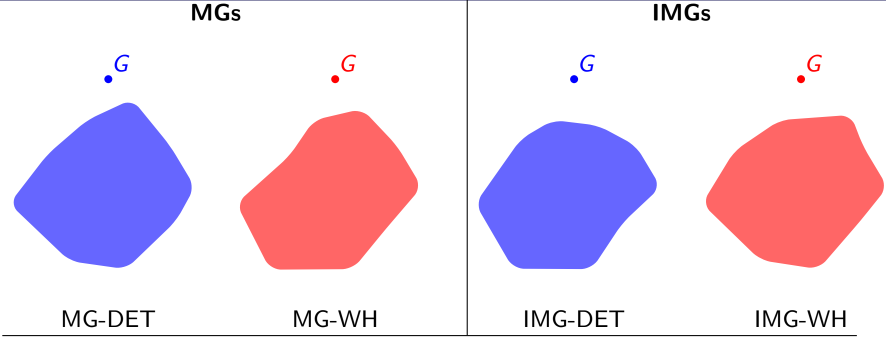
Training corpus: some combination of occurrences of the following.

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

- The learner knows correct analyses of wh-movement and reflexives.
- The learner must decide how to analyze 'foo': determiner or wh-phrase?

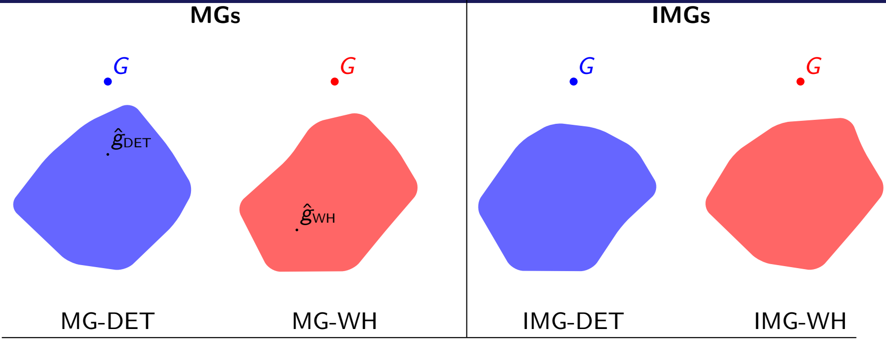



MGs
IMGs
G
G
G
G
MG-DET
MG-WH
IMG-DET
IMG-WH



Training corpus:

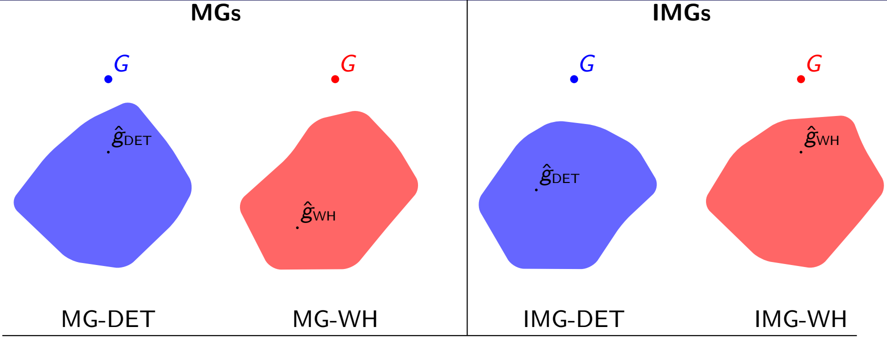
5	boys will shave
5	boys will shave themselves
5	who will shave
5	who will shave themselves
5	foo boys will shave



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

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

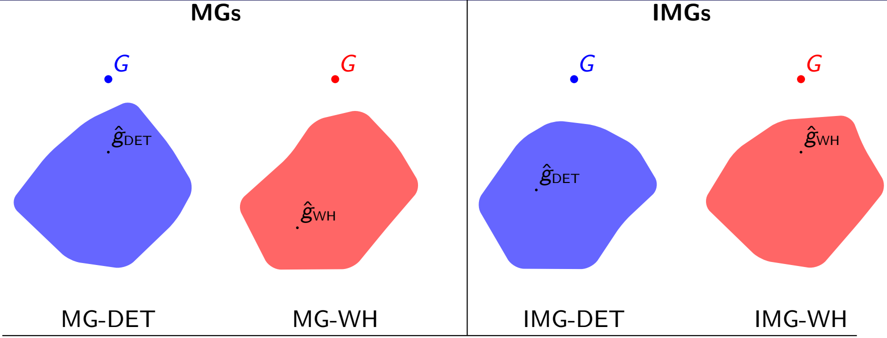


Training corpus:

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

$$\frac{P(D|\hat{g}_{\mathrm{DET}})}{P(D|\hat{g}_{\mathrm{WH}})} = \frac{3.36 \times 10^{-18}}{4.48 \times 10^{-20}} = 75.0$$

$$\frac{P(D|\hat{g}_{\mathrm{DET}})}{P(D|\hat{g}_{\mathrm{WH}})} = \frac{3.36 \times 10^{-18}}{2.45 \times 10^{-19}} = 13.7$$

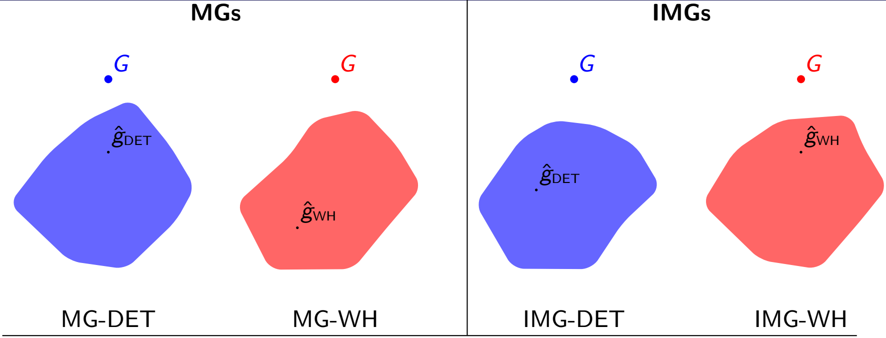


Training corpus:

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

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

$$\frac{P(D|\hat{g}_{\text{DET}})}{P(D|\hat{g}_{\text{WH}})} = \frac{7.64 \times 10^{-14}}{6.85 \times 10^{-10}} = 0.000112$$

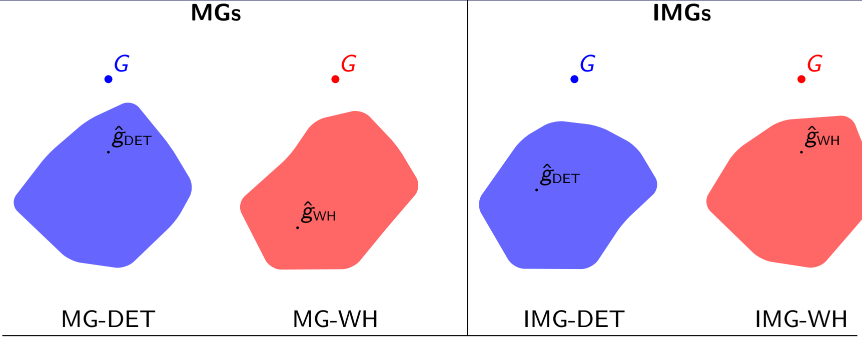


Training corpus:

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

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

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



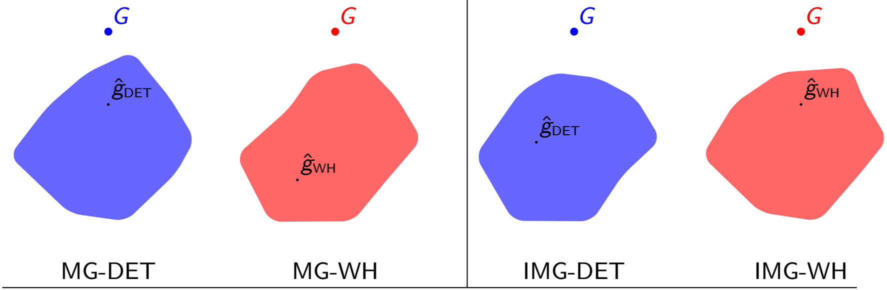
Training corpus:

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

$$\frac{P(D|\hat{g}_{\text{DET}})}{P(D|\hat{g}_{\text{WH}})} = \frac{2.83 \times 10^{-15}}{4.36 \times 10^{-20}} = 64900$$

$$\frac{P(D|\hat{g}_{\text{DET}})}{P(D|\hat{g}_{\text{WH}})} = \frac{1.31 \times 10^{-17}}{1.75 \times 10^{-17}} = 0.749$$

MGs | **IMGs**



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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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