Rebanking CCGbank for Improved NP Interpretation

Matthew Honnibal

June 23, 2010

Motivating Examples

CCG/HPSG/LFG/LTAG/etc promise semantic transparency:

Google acquired YouTube	$\rightarrow$	acquire(G, Y)	(1)
YouTube was acquired by Google	$\rightarrow$	acquire(G, Y)	(2)
Youtube, which Google acquired	$\rightarrow$	acquire(G, Y)	(3)
Google decided to acquire YouTube	$\rightarrow$	acquire(G, Y)	(4)

But what counts as surface variation?

Google's acquisition of YouTube	$\rightarrow$	?? (5)
Google's decision to acquire YouTube	$\rightarrow$	?? (6)
Google has to make a decision about acquiring YouTube	$\rightarrow$	?? (7)

Overview

- The project: Produce a CCG treebank that's as good as one we'd produce if we started from scratch.
- The process: Design analyses, then use existing Penn Treebank annotations to reform CCGbank.
- The point: Test semantic transparency against independent semantic annotations; use the corpus to improve CCG parsers and realisers.

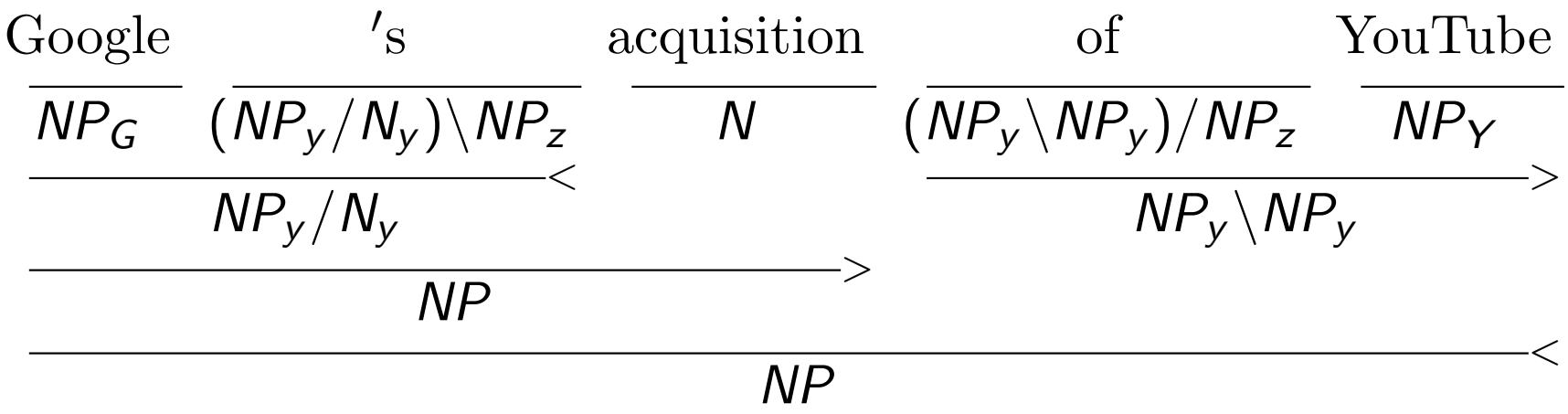
Semantic transparency in CCG

- Each word is assigned a lexical category and a logical form.
- Grammatical rules may only concatenate logical forms and bind variables.

$$\dfrac{\dfrac{\text{Google}}{NP_G} \quad \dfrac{\text{acquired}}{(S\backslash NP_y)/NP_z} \quad \dfrac{\text{YouTube}}{NP_Y}}{\dfrac{\text{acquire(y, z)}}{S\backslash NP_y}\,>}}{\dfrac{\text{acquire(y, YouTube)}}{S}\,<}$$

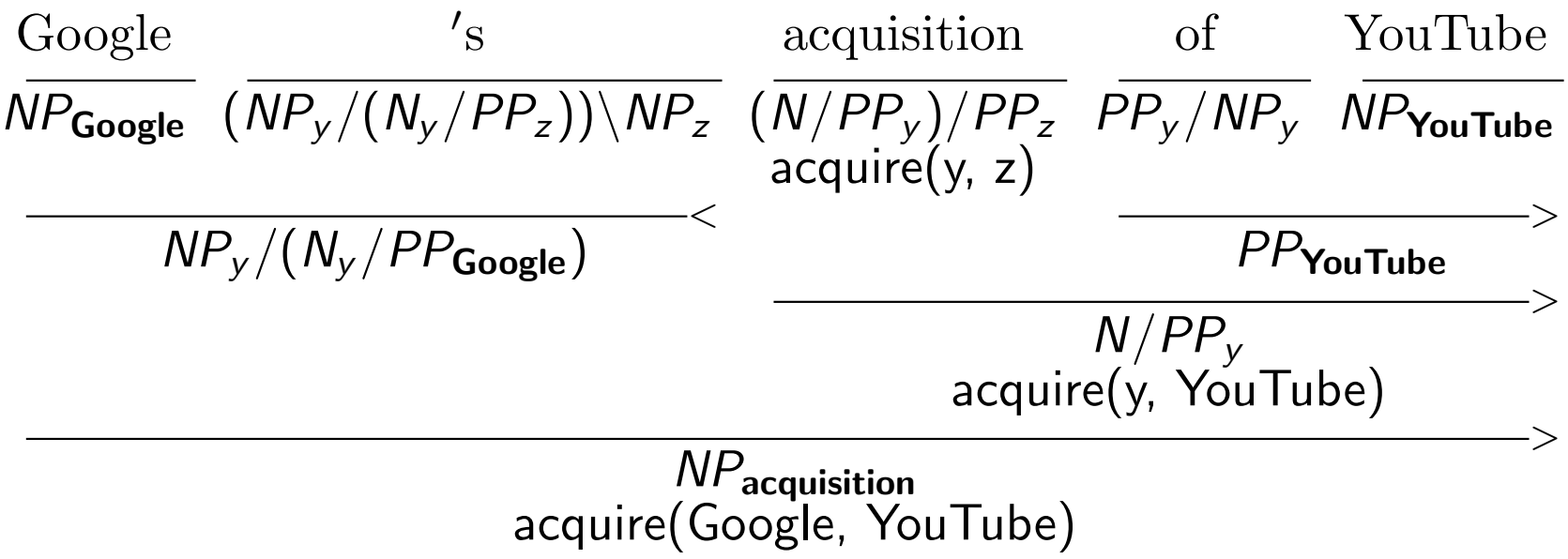
$$\text{acquire(Google, YouTube)}$$

Deverbal Nouns in CCGbank

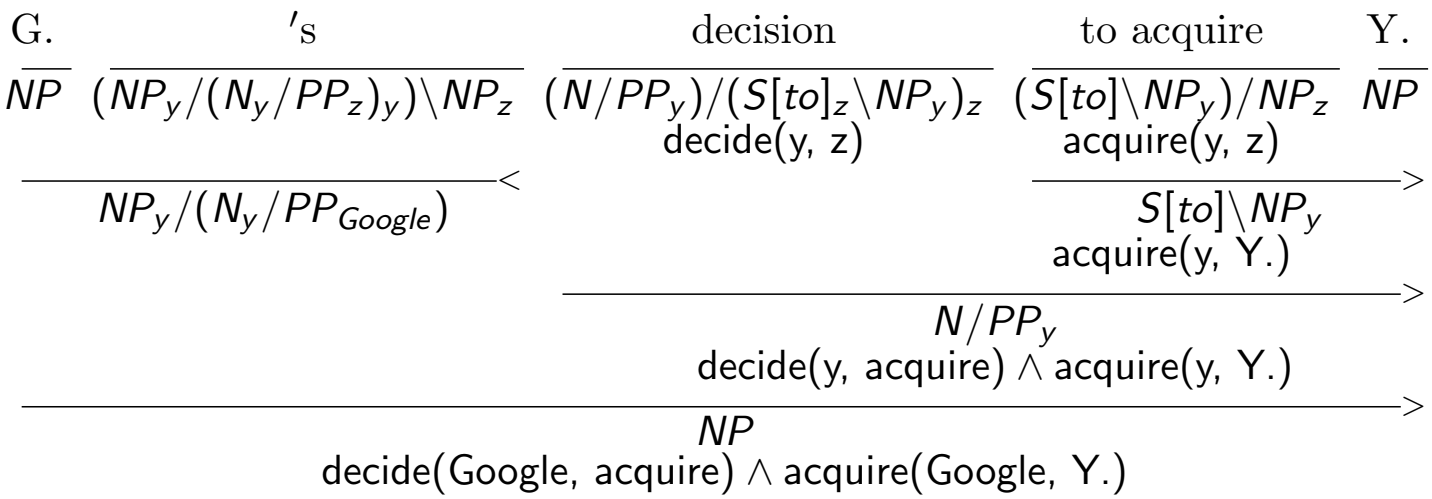


- CCGbank's analysis is derived from the Penn Treebank.
- Offers no support for nominal predicates: they do not subcategorise for their argument structures.

Our Analysis of Deverbal Nouns



Deverbal nouns passing long-range dependencies



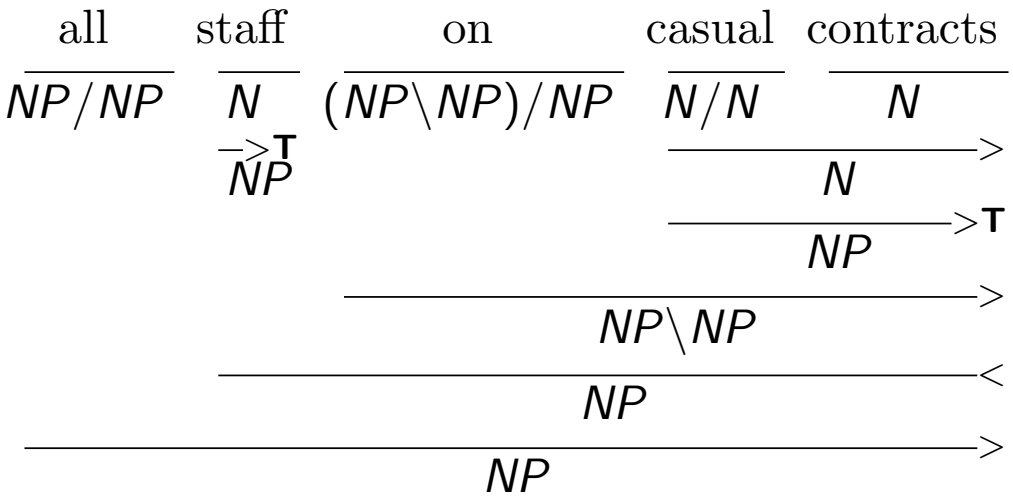
Implementation

- Implementation relied on NomBank, a predicate-argument annotation layer for the Penn Treebank
- Updating CCG trees intelligently is non-trivial, because label changes must be propagated down the tree
- 34,345 adnominal prepositional phrases converted to complements
- 18,919 left as adjuncts
- Most complementy preposition: *of*, 99.1% of occurrences as complement
- Most adjuncty preposition: *in*, 59.1% of occurrences as adjunct

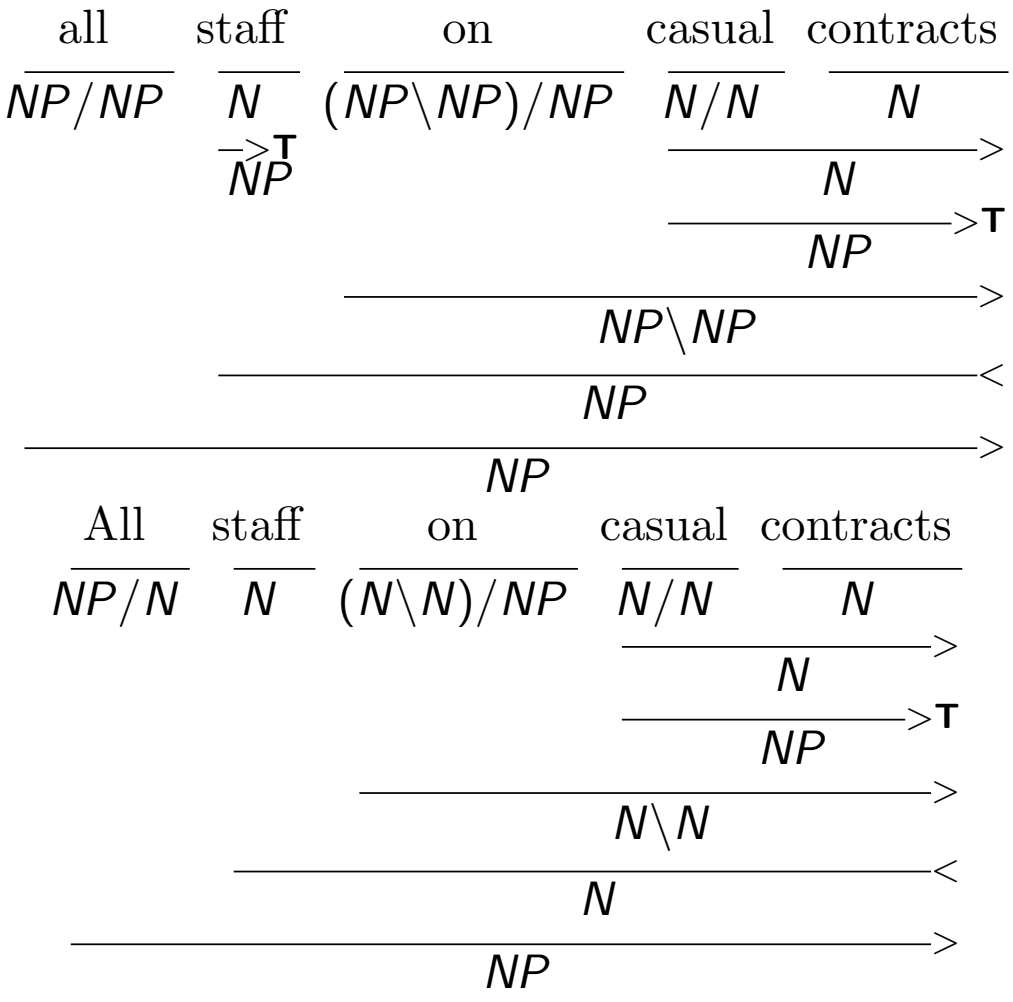
Merging Previous Changes

- NP bracketing from Dave Vadas's PhD
  e.g. Crude (oil prices) vs (Crude oil) prices
- Punctuation normalisation and quote restoration from Daniel Tse's honours
- Propbank complement/adjunct distinctions from a paper early in my PhD
- Verb particle constructions from James Constable's honours thesis,
  e.g. *he made up his mind* vs. *the parts are made up the river*

Restrictive vs. Non-restrictive Adnominals



Restrictive vs. Non-restrictive Adnominals



Partitive Constructions

$$\dfrac{\dfrac{\text{Four}}{NP}\quad\dfrac{\dfrac{\text{of}}{(NP_y\backslash NP_y)/NP_z}\quad\dfrac{\dfrac{\text{our}}{NP_y/N_y}\quad\dfrac{\text{members}}{N}}{NP_{\text{members}}}>}{(NP_y\backslash NP_y)_{\text{of}}}>}{NP_{\text{Four}}}<$$

Partitive Constructions

$$\dfrac{\dfrac{\text{Four}}{NP} \quad \dfrac{\dfrac{\text{of}}{(NP_y \backslash NP_y)/NP_z} \quad \dfrac{\dfrac{\text{our}}{NP_y/N_y} \quad \dfrac{\text{members}}{N}}{NP_{\text{members}}}>}{(NP_y \backslash NP_y)_{\text{of}}}>}{NP_{\text{Four}}}<$$

$$\dfrac{\dfrac{\text{Four}}{NP_y/PP_y} \quad \dfrac{\dfrac{\text{of}}{PP_y/NP_y} \quad \dfrac{\dfrac{\text{our}}{NP_y/N_y} \quad \dfrac{\text{members}}{N}}{NP_{\text{members}}}>}{PP_{\text{members}}}>}{NP_{\text{members}}}>$$

Extent of the Changes

Corpus	**L. Deps**	**U. Deps**	**Cats**
+NP brackets	97.2	97.7	98.5
+Quotes	97.2	97.7	98.5
+Propbank	93.0	94.9	96.7
+Particles	92.5	94.8	96.2
+Restrictivity	79.5	94.4	90.6
+Part. Gen.	76.1	90.1	90.4
+NP Pred-Arg	70.6	83.3	84.8

Table: Effect of the changes on CCGbank, by percentage of dependencies and categories left unchanged in Section 00.

Parsing Results

Corpus	WSJ 00			WSJ 23		
	lf	**uf**	**Cat**	**lf**	**uf**	**Cat**
CCGbank	87.2	92.9	94.1	87.7	93.0	94.4
+NP brackets	86.9	92.8	93.8	87.3	92.8	93.9
+Quotes	86.8	92.7	93.9	87.1	92.6	94.0
+Propbank	86.7	92.6	94.0	87.0	92.6	94.0
+Particles	86.4	92.5	93.8	86.8	92.6	93.8
+NP rebanking	84.2	91.2	91.9	84.7	91.3	92.2

Table: Parser evaluation on the rebanked corpora.

Intersection Evaluation

Corpus	Rebanked		CCGbank	
	lf	**uf**	**lf**	**uf**
+NP brackets	86.45	92.36	86.52	92.35
+Quotes	86.57	92.40	86.52	92.35
+Propbank	87.76	92.96	87.74	92.99
+Particles	87.50	92.77	87.67	92.93
+NP Rebanking	87.23	92.71	88.02	93.51

Table: Comparison of parsers trained on CCGbank and the rebanked corpora, using dependencies that occur in both.

Replacing the pipeline

- Research in computational linguistics is resource driven.
- The standard resource configuration suggests a pipeline: POS tagger $\rightarrow$ NE tagger $\rightarrow$ parser $\rightarrow$ semantic role labeller
- What we're trying to do is integrate the resources, and thereby integrate the tasks.
- Powerful linguistic representations (i.e. CCG) are central to our strategy.

Conclusion

- New and improved CCGbank including all previous updates
- Incorporation of SRL information from NomBank and PropBank into CCGbank
- New analyses to allow transparent interface between grammar and SRL-semantics
- 29.4% of dependencies in CCGbank updated
- Parsing remains feasible with the new, fine-grained resource