The Constituency of Hyperlinks in a Hypertext Corpus

mitcho (Michael Yoshitaka Erlewine)

Massachusetts Institute of Technology
mitcho@mitcho.com

International Society for the Linguistics of English
Boston University, June 19, 2011

The generative notion of constituency

- Certain substrings of sentences form natural units of linguistic import. Such units are called ***constituents***.
- Constituents are motivated and verified empirically by converging evidence of different kinds.

Constituency tests

(1) John ate an old hamburger.

Q: Is "an old hamburger" a constituent?

a) Clefting:
It's ***an old hamburger*** that John ate _____. *ok!*

b) Fronting:
An old hamburger, John ate _____, but a fresh orange, he didn't _____. *ok!*

c) Substitution:
Mary ate an old hamburger and John ate ***one*** too. *ok!*
("one" = "an old hamburger")

Constituency tests

(1) John ate an old hamburger.

Q: Is "ate an old" a constituent?

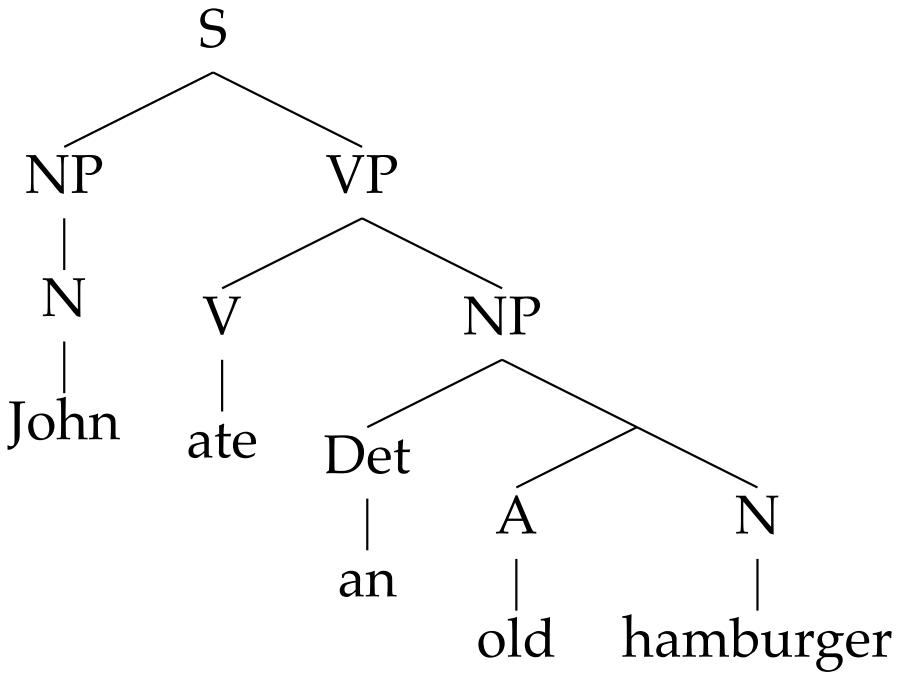
a) Clefting:
It's ***ate an old*** that John _____ hamburger. *no!*

b) Fronting:
Ate an old, John _____ hamburger... *no!*

c) Substitution:
Mary ate an old hamburger and John ***did*** sandwich too. *no!*
("did" ≠ "ate an old")

Constituency structure

Constituents are organized hierarchically, reflecting a phrase structure grammar:



Other converging evidence

- Other forms of converging evidence for constituency:
  - Pscholinguistic evidence (Fodor et al., 1974, a.o.)
  - Compositional semantics which tracks syntactic constituency (though perhaps not always perfectly), following Frege, Davidson, Montague

The limits of constituency tests

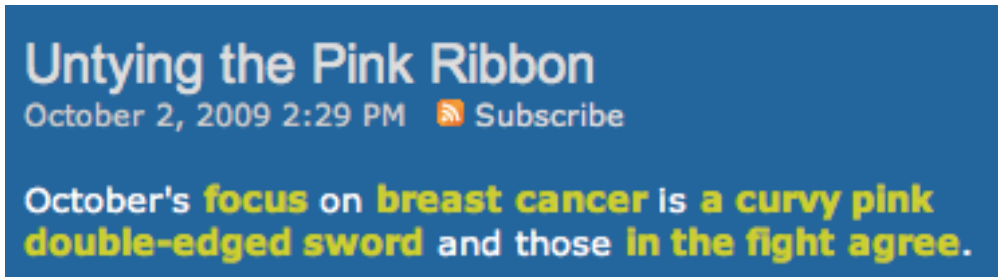
- Unfortunately, in some cases constituency tests may not apply or may yield conflicting results.
- Important proposals exist where constituency is at issue:
  - Binary branching (Kayne, 1984, a.o.)
    Branching in phrase structure grammars are always binary, not *n*-ary.
  - The DP hypothesis (Abney, 1987)
    D(eterminers) are the head of what have traditionally been labeled "Noun Phrases," with the D taking the Noun Phrase proper as its complement.
- As such, novel methodologies for constituency verification are welcome.

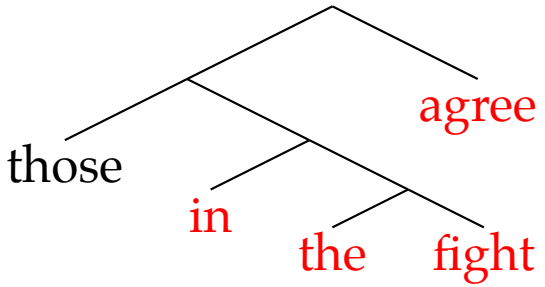
Hypertext and constituency

Observation:

- Not just *any* substring of sentences can be turned into ***hyperlinks***. Potential candidates seem to be rule-governed in some way.

`http://metafilter.com/85556`:





- The text "in the fight agree" is not a syntactic constituent.
- Upon closer inspection, it turns out this is actually two links:

(4) ... and those in the fight agree.

Goals

1. Test to what extent hyperlinks reflect the constituent structure of their host sentences.

   ☞ *Strong correlation!*

2. Present a novel class of linguistic data, non-constituent links, for further study.

A common insight: Spitovsky et al. (2010)

- A connection between HTML markup and dependencies
- Unsupervised grammar induction of a dependency-based parser (Klein and Manning, 2004) on a hypertext corpus, with constraints limiting dependencies from within each markup region
- 5% improvement over previous state-of-the-art
- But only minimal discussion of what kinds of linguistic objects hyperlinks are

Methodology

Corpus:

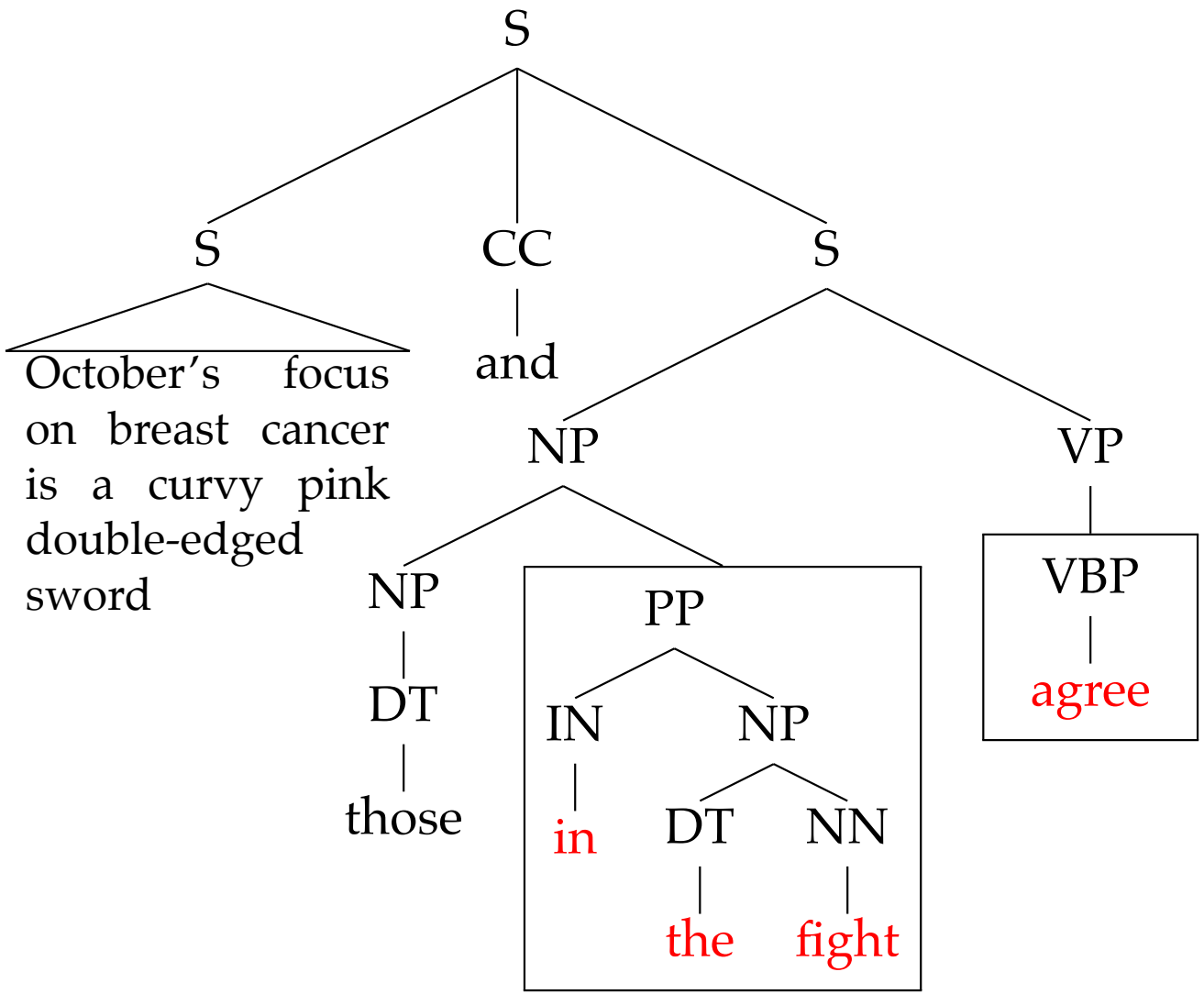
- MetaFilter (`http://metafilter.com`), a large, link-rich website. Currently about 100,000 "entries."
- 5.7m words, 375k human-annotated links.

Evaluation:

- Statistical parsing in lieu of manual coding, as a first approximation
- Parse the entry texts using the Stanford Parser (Klein and Manning, 2003) trained primarily on the Wall Street Journal section of the Penn Treebank (PTB; Marcus 1993).
- Find the subset of the parse tree that corresponds to the link.
- Check if this is a constituent.

Methodology

Entry 85556:



Results

A work-in-progress metric:

76.2% of all hyperlinks in the corpus are constituents.

- This value is after one type semi-supervised correction of noun phrase structure.
- "Out of the box": 72%
- Choosing random subsentences (null hypothesis) we would expect ≈27.6% constituency.
- Preliminary sampling and manual coding indicates an overwhelming number of false negatives.

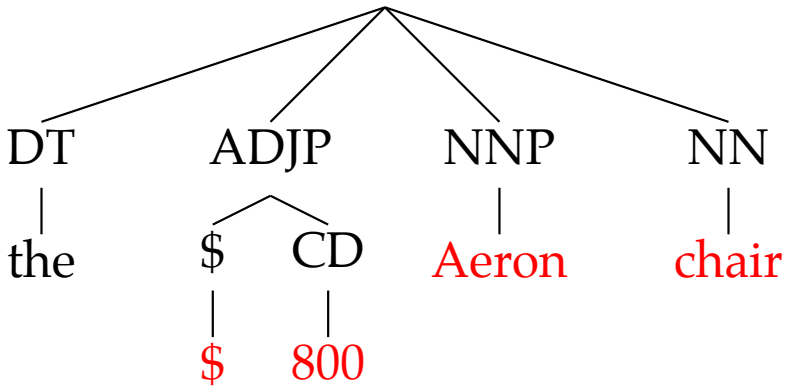
Average number of words per sentence: 15.658 (≈ 16)
P(link being constituent in 15-word sentence) =

$$\frac{\text{constituents in 15-word sentence}}{\text{number of subsentences}} = \frac{15+15-1}{\binom{15}{2}} = \frac{29}{105} = 27.6\%$$

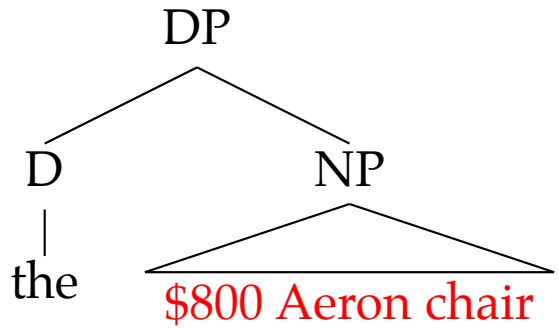
Sources of error: *n*-ary branching

- The Stanford Parser trained on the PTB produces *n*-ary branching structures (5a).
- A common configuration tagged by this methodology as a "non-constituent" are noun phrases missing their Determiners.

(5) a.



b.



- In a modern syntax following Abney's (1987) DP hypothesis, "$800 Aeron chair" would actually be a constituent (5b).
- This source of error has been adjusted for.

Types of links by POS

Lowest node dominating all of the link:

POS	N	%
NP	150458	39.9986
S	46434	12.3443
NNP	30651	8.1484
VP	25487	6.7756
NN	25173	6.6921
NNS	12739	3.3866
JJ	11228	2.9849
RB	7703	2.0478
CD	7201	1.9144
PRN	6527	1.7352
FRAG	5409	1.4380
PP	4312	1.1463
...		<1

- Over 58% nominal
- Spitovsky et al. (2010) found 74.5% to be nominal using the same metric, but with a different corpus.
- 12.3% sentential, 6.8% verb phrase-level

A typology of "non-constituents"

- Links deemed to be "non-constituents" by this methodology are then categorized in terms of what material is missing which, if included, would result in a constituent.

(6) A Virginia jury has [found Ahmed Omar Abu Ali [guilty of terrorism related crimes]]. 46912

⇒ Missing: PP after the link

A typology of "non-constituent links"

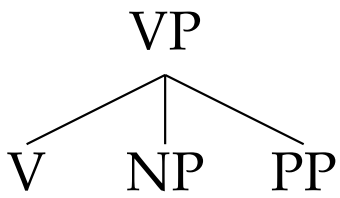
Missing nodes from links classified as "non-constituents":

category	position	N	%
PP	after	9166	12.17%
DT	before	8850	11.75%
NP	after	6173	8.19%
PRN	after	4834	6.42%
SBAR	after	4571	6.07%
JJ	before	4118	5.47%
NNP	after	3602	4.78%
NN	before	3286	4.36%
CC	after	2999	3.98%
NNP	before	2963	3.93%
VP	after	2859	3.79%
...			

- But it cannot just be that certain linguistic units in certain positions (PPs on the right) tend to be left off...

Grammatical sensitivity

- Consider the frame "V NP PP."
- If V transitive, PP adjunct. If V ditransitive, PP argument.
- Identical structure via the Stanford/PTB parser:

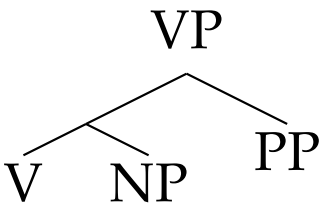


- The *n*-ary branching structure may again lead to false negatives.

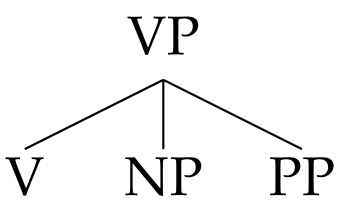
"V NP PP"

- A more modern syntactic theory would structurally distinguish the two PPs:[1]

PP adjunct:



PP argument of V:

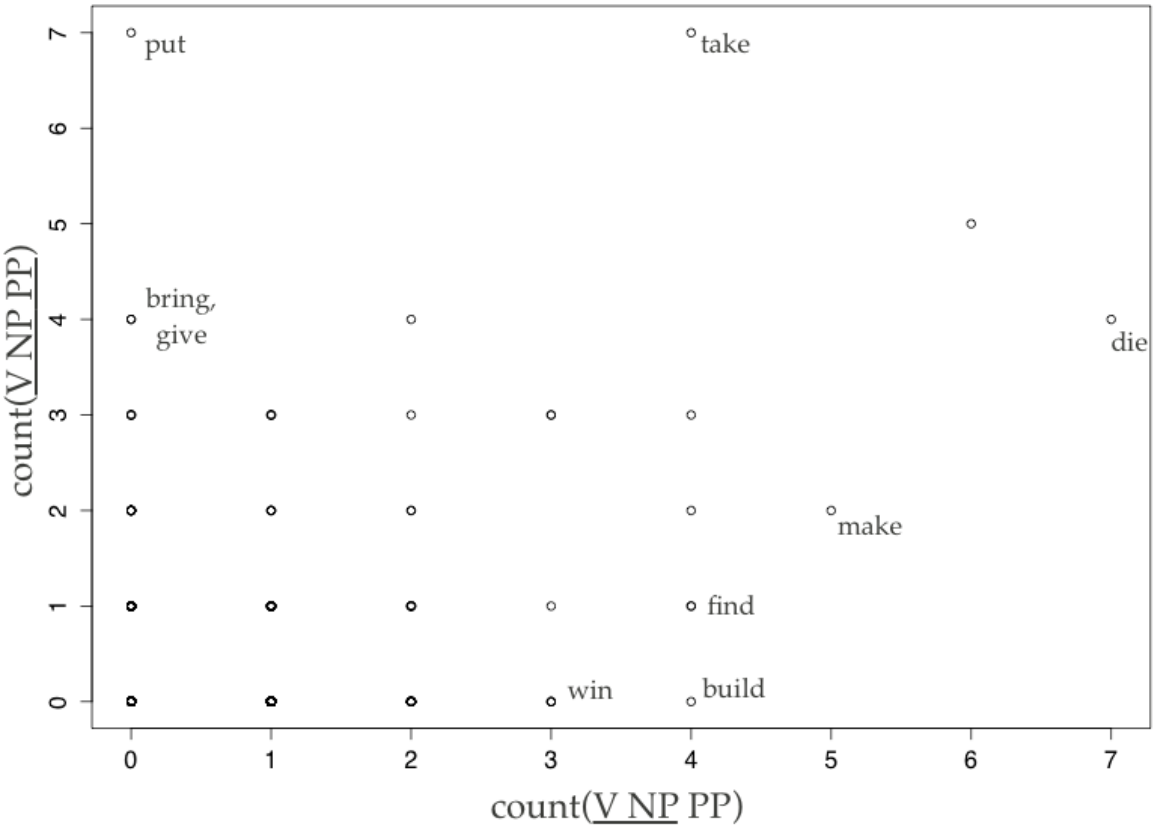


Predictions:

1. Linking "V NP PP" should entail that the PP is an adjunct.
2. Linking "V NP PP" but not "V NP PP" indicates that the V is ditransitive.

"V NP PP"

Verbs in "V NP PP" frame by link configurations



Non-constituent links

- Legitimate non-constituent links exist:

(7) ...the NY Times reports that the F.D.A. is cracking down. 21196

(8) If you're going to kill off an entire section of a newspaper and fire all of the staffers who work there, it's probably a good idea to get the Twitter password first. 87944

- Intuitively:

(8′) * ...kill off an entire section of a newspaper and fire all of the staffers who work there...

(8″) * ...kill off an entire section of a newspaper and fire all of the staffers who work there...

Non-constituent links

- These non-constituent links are not random; they are also rule-governed in some way.
  - Perhaps it's a semantic condition of referentiality?
  - The same string potentially being a constituent in a similar sentence?
- The legitimate non-constituent links seem to be an interesting class of linguistic objects which warrant further study.

Conclusion

1. Hyperlinks have a strong tendency to reflect the constituent structure of their host sentences, showing sensitivity to structural distinctions.
2. The seeds of a novel methodology of studying hyperlink markup in a hypertext corpus to investigate syntactic constituency.
3. True non-constituent links exist, but seem to form an interesting class of linguistic objects which warrants further study.
   - A novel type of linguistic data: a natural class of non-constituents

Next steps

- A more precise evaluation of the hyperlink-constituency hypothesis, using sampling and manual coding.
- Improvement of project corpus and tools, to be made publicly-accessible.
- Potentially, expansion of corpus and tools to another language.

Acknowledgements

Many thanks to my UROP researchers and contributors:

- Patrick Hulin, Patrick Hurst, and Antony Nguyen (MIT)
- Vedrana Janković (Faculty of Electrical Engineering and Computing, Croatia)

and to David Barner, David Pesetsky, and Stuart Shieber for comments and discussion. All errors are my own.

Abney, Steven. 1987. The English noun phrase in its sentential aspect. Doctoral Dissertation, Massachusetts Institute of Technology.

Fodor, Jerry Alan, Thomas G. Bever, and Merrill F. Garrett. 1974. *The psychology of language: an introduction to psycholinguistics and generative grammar*. McGraw-Hill.

Kayne, Richard. 1984. *Connectedness and binary branching*. Foris, Dordrecht.

Klein, Dan, and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 423–430.

Klein, Dan, and Christopher D. Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *ACL*.

Larson, Richard K. 1988. On the double object construction. *Linguistic Inquiry* 29:335–392.

Marcus, Mitchell P. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* 19.

Spitovsky, Valentin I., Daniel Jurafsky, and Hiyan Alshawi. 2010. Profiting from mark-up: Hyper-text annotations for guided parsing. In *Proceedings of ACL-2010*.