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
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.
(1) John ate an old hamburger.

Q: Is “an old hamburger” a constituent?

a) Clefting:
   It’s an old hamburger that John ate _______. \(\text{ok!}\)

b) Fronting:
   An old hamburger, John ate _______, but a fresh orange, he didn’t _______. \(\text{ok!}\)

c) Substitution:
   Mary ate an old hamburger and John ate \textit{one} too. \(\text{ok!}\)
   (“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")
Constituents are organized hierarchically, reflecting a phrase structure grammar:

```
S
  | NP
  |   | V
  |   | NP
  |   |   | Det
  |   |   | A
  |   |   | N
  |   | ate
  |   | an
  | John
  | old
  | hamburger
```
Other converging evidence

- Other forms of converging evidence for constituency:
  - Psycholinguistic 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:

October’s focus on breast cancer is a curvy pink double-edged sword.
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 \( \approx 27.6\% \) constituency.
- Preliminary sampling and manual coding indicates an overwhelming number of false negatives.

Average number of words per sentence: 15.658 (\( \approx 16 \))

\[
P(\text{link being constituent in 15-word sentence}) = \frac{\text{constituents in 15-word sentence}}{\text{number of subsentences}} = \frac{15 + 15 - 1}{\binom{15}{2}} = \frac{29}{105} = 27.6\%
\]
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. NP
  - DT
  - ADJP
  - NNP: Aeron
  - NN: chair

- b. DP
  - D
  - NP: $800 Aeron chair

- 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:

<table>
<thead>
<tr>
<th>POS</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP</td>
<td>150458</td>
<td>39.9986</td>
</tr>
<tr>
<td>S</td>
<td>46434</td>
<td>12.3443</td>
</tr>
<tr>
<td>NNP</td>
<td>30651</td>
<td>8.1484</td>
</tr>
<tr>
<td>VP</td>
<td>25487</td>
<td>6.7756</td>
</tr>
<tr>
<td>NN</td>
<td>25173</td>
<td>6.6921</td>
</tr>
<tr>
<td>NNS</td>
<td>12739</td>
<td>3.3866</td>
</tr>
<tr>
<td>JJ</td>
<td>11228</td>
<td>2.9849</td>
</tr>
<tr>
<td>RB</td>
<td>7703</td>
<td>2.0478</td>
</tr>
<tr>
<td>CD</td>
<td>7201</td>
<td>1.9144</td>
</tr>
<tr>
<td>PRN</td>
<td>6527</td>
<td>1.7352</td>
</tr>
<tr>
<td>FRAG</td>
<td>5409</td>
<td>1.4380</td>
</tr>
<tr>
<td>PP</td>
<td>4312</td>
<td>1.1463</td>
</tr>
<tr>
<td>...</td>
<td>&lt;1</td>
<td></td>
</tr>
</tbody>
</table>

- 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]].

⇒ Missing: PP after the link
A typology of “non-constituent links”

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

But it cannot just be that certain linguistic units in certain positions (PPs on the right) tend to be left off...
Consider the frame “V NP PP.”

If V transitive, PP adjunct. If V ditransitive, PP argument.

Identical structure via the Stanford/PTB parser:

```
VP
  V
  NP
  PP
```

The $n$-ary branching structure may again lead to false negatives.
A more modern syntactic theory would structurally distinguish the two PPs:

1. **PP adjunct:**
   - VP
   - V
   - NP
   - PP

2. **PP argument of V:**
   - VP
   - V
   - NP
   - PP

**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.
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.
Hyperlinks have a strong tendency to reflect the constituent structure of their host sentences, showing sensitivity to structural distinctions.

The seeds of a novel methodology of studying hyperlink markup in a hypertext corpus to investigate syntactic constituency.

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.


