NLP
Introduction to NLP

The Penn Treebank
Description

- **Background**
  - From the early ‘90s
  - Developed at the University of Pennsylvania
  - (Marcus, Santorini, and Marcinkiewicz 1993)
- **Size**
  - 40,000 training sentences
  - 2,400 test sentences
- **Genre**
  - Mostly Wall Street Journal news stories and some spoken conversations
- **Importance**
  - Helped launch modern automatic parsing methods
External links

• Treebank−3
  – http://catalog.ldc.upenn.edu/LDC99T42

• Original version
  – http://catalog.ldc.upenn.edu/LDC95T7

• Tokenization guidelines
  – http://www.cis.upenn.edu/~treebank/tokenization.html

• The American National Corpus
  – http://www.anc.org
<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>coordinating conjunction</td>
<td>and</td>
</tr>
<tr>
<td>CD</td>
<td>cardinal number</td>
<td>1, third</td>
</tr>
<tr>
<td>DT</td>
<td>determiner</td>
<td>the</td>
</tr>
<tr>
<td>EX</td>
<td>existential there</td>
<td>there is</td>
</tr>
<tr>
<td>FW</td>
<td>foreign word</td>
<td>d'oeuvre</td>
</tr>
<tr>
<td>IN</td>
<td>preposition/subordinating conjunction</td>
<td>in, of, like</td>
</tr>
<tr>
<td>JJ</td>
<td>adjective</td>
<td>green</td>
</tr>
<tr>
<td>JJR</td>
<td>adjective, comparative</td>
<td>greener</td>
</tr>
<tr>
<td>JJS</td>
<td>adjective, superlative</td>
<td>greenest</td>
</tr>
<tr>
<td>LS</td>
<td>list marker</td>
<td>1)</td>
</tr>
<tr>
<td>MD</td>
<td>modal</td>
<td>could, will</td>
</tr>
<tr>
<td>NN</td>
<td>noun, singular or mass</td>
<td>table</td>
</tr>
<tr>
<td>NNS</td>
<td>noun plural</td>
<td>tables</td>
</tr>
<tr>
<td>NNP</td>
<td>proper noun, singular</td>
<td>John</td>
</tr>
<tr>
<td>NNPS</td>
<td>proper noun, plural</td>
<td>Vikings</td>
</tr>
<tr>
<td>PDT</td>
<td>predeterminer</td>
<td>both the boys</td>
</tr>
<tr>
<td>POS</td>
<td>possessive ending</td>
<td>friend’s</td>
</tr>
</tbody>
</table>
# Penn Treebank Tagset (2/2)

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRP</td>
<td>personal pronoun</td>
<td>I, he, it</td>
</tr>
<tr>
<td>PRP$</td>
<td>possessive pronoun</td>
<td>my, his</td>
</tr>
<tr>
<td>RB</td>
<td>adverb</td>
<td>however, usually, naturally, here, good</td>
</tr>
<tr>
<td>RBR</td>
<td>adverb, comparative</td>
<td>better</td>
</tr>
<tr>
<td>RBS</td>
<td>adverb, superlative</td>
<td>best</td>
</tr>
<tr>
<td>RP</td>
<td>particle</td>
<td>give up</td>
</tr>
<tr>
<td>TO</td>
<td>to</td>
<td>to go, to him</td>
</tr>
<tr>
<td>UH</td>
<td>interjection</td>
<td>uhhuhhuhh</td>
</tr>
<tr>
<td>VB</td>
<td>verb, base form</td>
<td>take</td>
</tr>
<tr>
<td>VBD</td>
<td>verb, past tense</td>
<td>took</td>
</tr>
<tr>
<td>VBG</td>
<td>verb, gerund/present participle</td>
<td>taking</td>
</tr>
<tr>
<td>VBN</td>
<td>verb, past participle</td>
<td>taken</td>
</tr>
<tr>
<td>VBP</td>
<td>verb, sing. present, non-3d</td>
<td>take</td>
</tr>
<tr>
<td>VBZ</td>
<td>verb, 3rd person sing. present</td>
<td>takes</td>
</tr>
<tr>
<td>WDT</td>
<td>wh-determiner</td>
<td>which</td>
</tr>
<tr>
<td>WP</td>
<td>wh-pronoun</td>
<td>who, what</td>
</tr>
<tr>
<td>WPS</td>
<td>possessive wh-pronoun</td>
<td>whose</td>
</tr>
<tr>
<td>WRB</td>
<td>wh-abverb</td>
<td>where, when</td>
</tr>
</tbody>
</table>
Because the CD had an effective yield of 13.4% when it was issued in 1984, and interest rates in general had declined sharply since then, part of the price Dr. Blumenfeld paid was a premium -- an additional amount on top of the CD's base value plus accrued interest that represented the CD's increased market value.
Because the CD had an effective yield of 13.4%, it was issued in 1984.
(VP
  (VBD was)
  (NP-PRD
    (NP (DT a) (NN premium))
    (: --)
    (NP
      (NP (DT an) (JJ additional) (NN amount))
      (PP-LOC
        (IN on)
        (NP (NP (IN of) (NP (DT the) (NN price)) (SBAR (WHNP-3 (-NONE- 0))))
          (S
            (NP-SBJ (NNP Dr.) (NNP Blumenfeld))
            (VP (VBD paid) (NP (-NONE- *T*-3))))))))
  (CC plus)
  (NP (VBN accrued) (NN interest)))
(SBAR
  (WHNP-2 (WDT that))
  (S
    (NP-SBJ (-NONE- *T*-2))
    (VP
      (VBD represented)
      (NP
        (NP (DT the) (NNP CD) (POS 's))
        (VBN increased)
        (NN market)
        (NN value)))))

(., .)
Peculiarities

- **Complementizers**
  - e.g., “that”
- **Gaps**
  - *NONE*
- **SBAR**
  - SBAR → COMP S
  - E.g., “that *NONE* represented the CD’ market value”
tgrep

A < B  A immediately dominates B
A << B  A dominates B
A <- B  B is the last child of A
A <<, B  B is a leftmost descendant of A
A <<` B  B is a rightmost descendant of A
A . B  A immediately precedes B
A .. B  A precedes B
A $ B  A and B are sisters
A $. B  A and B are sisters and A immediately precedes B
A $.. B  A and B are sisters and A precedes B
The use of treebanks

• Disadvantages
  – A lot more work to annotate 40K+ sentences than to write a grammar.

• Advantages
  – Statistics about different constituents and phenomena
  – Training systems
  – Evaluating systems
  – Multilingual extensions
Introduction to NLP

Issues with Context-free Grammars
Agreement

- **Number**
  - Chen is/people are

- **Person**
  - I am/Chen is

- **Tense**
  - Chen was reading/Chen is reading/Chen will be reading

- **Case**
  - not in English but in many other languages such as German, Russian, Greek

- **Gender**
  - not in English but in many other languages such as German, French, Spanish
Combinatorial explosion

• Many combinations of rules are needed to express agreement
  – $S \rightarrow NP \ VP$
  – $S \rightarrow 1sgNP \ 1sgVP$
  – $S \rightarrow 2sgNP \ 2sgVP$
  – $S \rightarrow 3sgNP \ 3sgVP$
  – ...
  – $1sgNP \rightarrow 1sgN$
  – ...
Subcategorization frames

- Direct object
  - The dog ate a sausage

- Prepositional phrase
  - Mary left the car in the garage

- Predicative adjective
  - The receptionist looked worried

- Bare infinitive
  - She helped me buy this place

- To-infinitive
  - The girl wanted to be alone

- Participial phrase
  - He stayed crying after the movie ended

- That-clause
  - Ravi doesn’t believe that it will rain tomorrow

- Question-form clauses
  - She wondered where to go
CFG independence assumptions

- Non-independence
  - All NPs
    - 11% NP PP, 9% DT NN, 6% PRP
  - NPs under S
    - 9% NP PP, 9% DT NN, 21% PRP
  - NPs under VP
    - 23% NP PP, 7% DT NN, 4% PRP
  - (example from Dan Klein)

- Lexicalized grammars
  - later
Conclusions

• Syntax helps understand the meaning of a sentence.
  – Bob gave Alice a flower
  – Who gave a flower to Alice?
  – What did Bob give to Alice?

• Context-free grammars are an appropriate representation for syntactic information

• Dynamic programming is needed for efficient parsing
  – Cubic time to find one parse
  – Still exponential time to find all parses
  – Why?
Answer

• Why does it still take an exponential time to find all parses?
  – Very simple – because the number of parses can be exponential
NLP