Parsing
What is Parsing?

S → NP VP
NP → Det N
NP → NP PP
VP → V NP
VP → VP PP
PP → P NP

NP → Papa
N → caviar
N → spoon
V → spoon
V → ate
P → with
Det → the
Det → a

Papa ate the caviar with a spoon
What is Parsing?

S → NP VP
NP → Det N
NP → NP PP
VP → V NP
VP → VP PP
PP → P NP

S
\[ \overset{\text{NP}}{\text{NP}} \quad \overset{\text{VP}}{\text{VP}} \]
\[ \overset{\text{Papa}}{\text{VP}} \quad \overset{\text{ate}}{\text{VP}} \]
\[ \overset{\text{the caviar}}{\text{PP}} \quad \overset{\text{a spoon}}{\text{PP}} \]
Programming languages

printf ("/charset [%s",
   (re_opcode_t) *(p - 1) == charset_not ? "^" : "") ;
assert (p + *p < pend);
for (c = 0; c < 256; c++)
   if (c / 8 < *p && (p[1 + (c/8)] & (1 << (c % 8)))) {
      /* Are we starting a range? */
      if (last + 1 == c && ! inrange) {
         putchar ('-');
         inrange = 1;
      }
      /* Have we broken a range? */
      else if (last + 1 != c && ! inrange) {
         putchar (last);
         inrange = 0;
      }
   }
   if (! inrange)
      putchar (c);
   last = c;

- Easy to parse.
- Designed that way!
Natural languages

- No {} () [] to indicate scope & precedence
- Lots of overloading (arity varies)
- Grammar isn’t known in advance!
- Context-free grammar not best formalism
Ambiguity

S → NP VP
NP → Det N
NP → NP PP
VP → V NP
VP → VP PP
PP → P NP

S

NP

Papa

VP

ate

Det

N

with

Det

N

PP

the caviar

a spoon
Ambiguity

S → NP VP
NP → Det N
NP → NP PP
VP → V NP
VP → VP PP
PP → P NP

S
  \|-- NP
  \  \-- V
    \  \-- NP
      \  \-- PP
        \  \-- Det
          \  \-- N
            \  \-- Det
              \  \-- N
                \  \-- P
                  \  \-- Det
                    \  \-- N
                      \  \-- a
                        \-- spoon

NP → Papa
N → caviar
N → spoon
V → spoon
V → ate
P → with
Det → the
Det → a
The parsing problem

Recent parsers quite accurate
... good enough to help a range of NLP tasks!
Applications of parsing (1/2)

- **Machine translation** (Alshawi 1996, Wu 1997, ...)
  
  ![Diagram of tree operations]

- **Speech synthesis from parses** (Prevost 1996)
  
  The government plans to raise income tax.
  The government plans to raise income tax the imagination.

- **Speech recognition using parsing** (Chelba et al 1998)
  
  Put the file in the folder.
  Put the file and the folder.
Applications of parsing (2/2)

- Grammar checking (Microsoft)
- Indexing for information retrieval (Woods 1997)
  ... washing a car with a hose ... vehicle maintenance
- Information extraction (Hobbs 1996)

Warning: these slides are out of date
Most linguistic properties are defined over trees.

One needs to parse to see subtle distinctions. E.g.:

Sara dislikes criticism of her. \( (\text{her} \neq \text{Sara}) \)
Sara dislikes criticism of her by anyone. \( (\text{her} \neq \text{Sara}) \)
Sara dislikes anyone’s criticism of her. \( (\text{her} = \text{Sara} \text{ or } \text{her} \neq \text{Sara}) \)
What is meaning of $3 + 5 \times 6$?

First parse it into $3 + (5 \times 6)$
What is meaning of $3 + 5 \times 6$?

First parse it into $3 + (5 \times 6)$

Now give a meaning to each node in the tree (bottom-up)
assert(every(nation, λx ∃e present(e), act(e,wanting), wanter(e,x), wantee(e, λe’ act(e’,loving), lover(e’,G), lovee(e’,L))))

every every nation
λv λx ∃e present(e), v(x)(e)
λy λx λe act(e,wanting), wanter(e,x), wantee(e,y)
λy λx λe act(e,loving), lover(e,x), lovee(e,y)
Now let’s develop some parsing algorithms!

What substrings of this sentence are NPs?

Which substrings could be NPs in the right context?

Which substrings could be other types of phrases?

How would you defend a substring on your list?
“Papa ate the caviar with a spoon”
First try ... does it work?

- for each constituent on the LIST (Y I Mid)
  - scan the LIST for an adjacent constituent (Z Mid J)
  - if grammar has a rule to combine them (X \rightarrow Y Z)
    - then add the result to the LIST (X I J)

"Papa ate the caviar with a spoon"
“Papa ate the caviar with a spoon”

Second try …

- initialize the list with parts-of-speech \((T J-1 J)\)
  where \(T\) is a preterminal tag (like Noun) for the \(J^{th}\) word

- for each constituent on the LIST \((Y I Mid)\)
  - scan the LIST for an adjacent constituent \((Z Mid J)\)
  - if grammar has a rule to combine them \((X \rightarrow Y Z)\)
    - then add the result to the LIST \((X I J)\)

- if the above loop added anything, do it again!
  (so that \(X I J\) gets a chance to combine or be combined with)
Third try …

- Initialize the list with parts-of-speech \((T \ J-1 \ J)\)
  where \(T\) is a preterminal tag (like Noun) for the \(J^{th}\) word

- For each constituent on the LIST \((Y \ I \ Mid)\)
  - For each adjacent constituent on the list \((Z \ Mid \ J)\)
    - For each rule to combine them \((X \rightarrow Y \ Z)\)
      - Add the result to the LIST \((X \ I \ J)\)
        if it’s not already there

- If the above loop added anything, do it again!
  (so that \(X \ I \ J\) gets a chance to combine or be combined with)
Third try …

<table>
<thead>
<tr>
<th>Initialize</th>
<th>1st pass</th>
<th>2nd pass</th>
<th>3rd pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP 0 1</td>
<td>NP 2 4</td>
<td>VP 1 4</td>
<td>...</td>
</tr>
<tr>
<td>V 1 2</td>
<td>NP 5 7</td>
<td>NP 2 4</td>
<td>...</td>
</tr>
<tr>
<td>Det 2 3</td>
<td>PP 4 7</td>
<td>NP 5 7</td>
<td></td>
</tr>
<tr>
<td>N 3 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P 4 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Det 5 6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N 6 7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V 6 7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
"Papa ate the caviar with a spoon"

Follow backpointers to get the parse.
“Papa ate the caviar with a spoon”
Correct but still inefficient ...

We kept checking the same pairs that we’d checked before (both bad and good pairs)

Can’t we manage the process in a way that avoids duplicate work?
Correct but still inefficient ...

We kept checking the same pairs that we’d checked before (both bad and good pairs)

Can’t we manage the process in a way that avoids duplicate work?

And even finding new pairs was expensive because we had to scan the whole list

Can’t we have some kind of index that will help us find adjacent pairs?
Indexing: Store S 0 4 in chart[0,4]
Avoid duplicate work:
Build width-1, then width-2, etc.
How do we build a width-6 phrase?
(after building and indexing all shorter phrases)

Papa ate the caviar with a spoon

<table>
<thead>
<tr>
<th>start position</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>NP</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
</tr>
<tr>
<td>1</td>
<td>V</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
</tr>
<tr>
<td>2</td>
<td>Det</td>
<td>NP</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
</tr>
<tr>
<td>4</td>
<td>P</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
</tr>
<tr>
<td>5</td>
<td>Det</td>
<td>NP</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
</tr>
<tr>
<td>6</td>
<td>N</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
</tr>
<tr>
<td>7</td>
<td>V</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
<td>⦸</td>
</tr>
</tbody>
</table>

1 7 = 12 + 2 7 ☺
1 3 + 3 7
1 4 + 4 7 ☻
1 5 + 5 7
1 6 + 6 7
Avoid duplicate work:
Build width-1, then width-2, etc.
CKY algorithm, recognizer version

- **Input**: string of n words
- **Output**: yes/no (since it’s only a recognizer)
- **Data structure**: n x n table
  - rows labeled 0 to n-1
  - columns labeled 1 to n
  - cell [i,j] lists constituents found between i and j

- Basic idea: fill in width-1 cells, then width-2, …
CKY algorithm, recognizer version

- for \( J := 1 \) to \( n \)
  - Add to \([J-1,J]\) all categories for the \( J^{th} \) word
- for width := 2 to \( n \)
  - for start := 0 to \( n-width \)  
    - Define end := start + width  
    - for mid := start+1 to end-1  
      - for every nonterminal \( Y \) in \([start,mid]\)  
        - for every nonterminal \( Z \) in \([mid,end]\)  
          - for all nonterminals \( X \)  
            - if \( X \rightarrow Y \) \( Z \) is in the grammar  
              - then add \( X \) to \([start,end]\)
Loose ends to tie up (no slides)

- What’s the space?
  - Duplicate entries?
  - How many parses could there be?
  - Can we fix this?

- What’s the runtime?
CKY algorithm, recognizer version

- for J := 1 to n
  - Add to [J-1,J] all categories for the Jth word
- for width := 2 to n
  - for start := 0 to n-width // this is I
    - Define end := start + width // this is J
    - for mid := start+1 to end-1 // find all I-to-J phrases
      - for every nonterminal Y in [start,mid]
        - for every nonterminal Z in [mid,end]
          - for all nonterminals X
            - if X \rightarrow Y Z is in the grammar
              - then add X to [start,end]
Alternative version of inner loops

- for $J := 1$ to $n$
  - Add to $[J-1,J]$ all categories for the $J^{th}$ word
- for $\text{width} := 2$ to $n$
  - for $\text{start} := 0$ to $n-\text{width}$ // this is $I$
    - Define $\text{end} := \text{start} + \text{width}$ // this is $J$
    - for $\text{mid} := \text{start} + 1$ to $\text{end} - 1$ // find all $I$-to-$J$ phrases
      - for every rule $X \rightarrow Y Z$ in the grammar
        - if $Y$ in $[\text{start}, \text{mid}]$ and $Z$ in $[\text{mid}, \text{end}]$
          - then add $X$ to $[\text{start}, \text{end}]$
Extensions to discuss (no slides)

- In Chinese, no spaces between the words!? 
- Incremental (left-to-right) parsing?
- We assumed rules like $X \rightarrow Y Z$ or $X \rightarrow$ word.
  - Grammar in “Chomsky Normal Form.” What if it’s not?
- How do we choose among parses?
Chart Parsing in Dyna

\[
\text{phrase}(X,I,J) :- \text{rewrite}(X,W), \text{word}(W,I,J).
\]
\[
\text{phrase}(X,I,J) :- \text{rewrite}(X,Y,Z), \text{phrase}(Y,I,Mid), \text{phrase}(Z,Mid,J).
\]
\[
\text{goal} :- \text{phrase}(\text{start\_symbol}, 0, \text{sentence\_length}).
\]
Understanding the Key Rule

Substring from I to J could be a phrase of category X

\[ \text{phrase}(X,I,J) \ :- \ \text{rewrite}(X,Y,Z), \ \text{phrase}(Y,I,\text{Mid}), \ \text{phrase}(Z,\text{Mid},J). \]

e.g., \(\text{phrase}("\text{VP"},1,7)\) rewrite("\text{VP"},"V"","NP") phrase("V",1,2), phrase("NP",2,7)

it breaks up into adjacent substrings (from I to Mid and Mid to J) that could be phrases of categories Y and Z

("an X can be made of a Y next to a Z")
A word is a phrase” (if grammar allows)

\[
\text{phrase}(X,I,J) \leftarrow \text{rewrite}(X,W), \text{word}(W,I,J).
\]

“Two adjacent phrases are a phrase” (if grammar allows)

\[
\text{phrase}(X,I,J) \leftarrow \text{rewrite}(X,Y,Z), \text{phrase}(Y,I,Mid), \text{phrase}(Z,Mid,J).
\]

“A phrase that covers the whole sentence is a parse”
(achieves our goal by showing that the sentence is grammatical)

\[
\text{goal} \leftarrow \text{phrase}(\text{start_symbol}, 0, \text{sentence_length}).
\]

Alternatively:

\[
\text{start_symbol} := “S”.
\text{sentence_length} := 7.
\]

sentence_length max= J for word(_,_,J).
Chart Parsing in Dyna

- We also need a sentence:

  word("Papa",0,1).
  word("ate",1,2).
  word("the",2,3).
  word("caviar",3,4).
  word("with",4,5).
  word("a",5,6).
  word("spoon",6,7).

- We also need a grammar:

  rewrite("NP","Papa").
  rewrite("N","caviar").
  rewrite("S","NP","VP").
  rewrite("NP","Det","N").
  rewrite("NP","NP","PP").
  ...
Procedural Algorithms

- The Dyna program runs fine.
- It nicely displays the abstract structure of the algorithm.

- But Dyna is a declarative programming language that hides the details of the actual execution from you.
- If you had to find the possible phrases by hand (or with a procedural programming language), what steps would you go through?
This picture assumes a slightly different version of the Dyna program, sorry.

**Discovered phrases & their relationships ("parse forest")**

When parsing the ambiguous sentence "Time flies like an arrow".
Discovered phrases & their relationships ("parse forest")

When parsing the ambiguous sentence "Time flies like an arrow"

This picture assumes a slightly different version of the Dyna program, sorry.
Procedural algorithms like CKY are just strategies for running this declarative program

\begin{verbatim}
phrase(X,I,J) :- rewrite(X,W), word(W,I,J).
phrase(X,I,J) :- rewrite(X,Y,Z), phrase(Y,I,Mid), phrase(Z,Mid,J).
goal :- phrase(start_symbol, 0, sentence_length).
\end{verbatim}

- And we’ll look at further such strategies later, e.g.
  - magic sets / Earley’s algorithm / left-to-right parsing,
  - agenda-based forward chaining,
  - backward chaining with memoization,
  - coarse-to-fine search,
  - probability-guided search and pruning, …