Phrase structure parsing organizes syntax into *constituents* or *brackets*.

- In general, this involves nested trees.
- Linguists can, and do, argue about details.
- Lots of ambiguity.
- Not the only kind of syntax...

```
new art critics write reviews with computers
```
Constituency Tests

- How do we know what nodes go in the tree?

- Classic constituency tests:
  - Substitution by *proform*
  - Question answers
  - Semantic grounds
    - Coherence
    - Reference
    - Idioms
  - Dislocation
  - Conjunction

- Cross-linguistic arguments, too
Conflicting Tests

- **Constituency isn’t always clear**
  - Units of transfer:
    - think about ~ penser à
    - talk about ~ hablar de
  - Phonological reduction:
    - I will go → I’ll go
    - I want to go → I wanna go
    - a le centre → au centre
  - Coordination
    - He went to and came from the store.

La vélocité des ondes sismiques
Classical NLP: Parsing

- Write symbolic or logical rules:
  
  Grammar (CFG)    | Lexicon
  
  ROOT → S    | NP → NP PP    | NN → interest
  S → NP VP  | VP → VBP NP   | NNS → raises
  NP → DT NN | VP → VBP NP PP | VBP → interest
  NP → NN NNS | PP → IN NP   | VBZ → raises
  ...

- Use deduction systems to prove parses from words
  - Minimal grammar on “Fed raises” sentence: 36 parses
  - Simple 10-rule grammar: 592 parses
  - Real-size grammar: many millions of parses

- This scaled very badly, didn’t yield broad-coverage tools
Ambiguities
The board approved [its acquisition] [by Royal Trustco Ltd.]
[of Toronto]
[for $27 a share]
[at its monthly meeting].
I cleaned the dishes from dinner

I cleaned the dishes with detergent

I cleaned the dishes in my pajamas

I cleaned the dishes in the sink
Syntactic Ambiguities I

- **Prepositional phrases:**
  \(\text{They cooked the beans in the pot on the stove with handles.}\)

- **Particle vs. preposition:**
  \(\text{The puppy tore up the staircase.}\)

- **Complement structures**
  \(\text{The tourists objected to the guide that they couldn’t hear.}\)
  \(\text{She knows you like the back of her hand.}\)

- **Gerund vs. participial adjective**
  \(\text{Visiting relatives can be boring.}\)
  \(\text{Changing schedules frequently confused passengers.}\)
Syntactic Ambiguities II

- **Modifier scope within NPs**
  
  *impractical design requirements*
  
  *plastic cup holder*

- **Multiple gap constructions**
  
  *The chicken is ready to eat.*
  
  *The contractors are rich enough to sue.*

- **Coordination scope:**
  
  *Small rats and mice can squeeze into holes or cracks in the wall.*
**Dark Ambiguities**

- **Dark ambiguities**: most analyses are shockingly bad (meaning, they don’t have an interpretation you can get your mind around)

This analysis corresponds to the correct parse of

“This will panic buyers! ”

- **Unknown words and new usages**
- **Solution**: We need mechanisms to focus attention on the best ones, probabilistic techniques do this
PCFGs
Probabilistic Context-Free Grammars

- A context-free grammar is a tuple \( <N, T, S, R> \)
  - \( N \) : the set of non-terminals
    - Phrasal categories: \( S, NP, VP, ADJP, \) etc.
    - Parts-of-speech (pre-terminals): \( NN, JJ, DT, VB \)
  - \( T \) : the set of terminals (the words)
  - \( S \) : the start symbol
    - Often written as ROOT or TOP
    - Not usually the sentence non-terminal \( S \)
  - \( R \) : the set of rules
    - Of the form \( X \rightarrow Y_1 Y_2 \ldots Y_k \), with \( X, Y_i \in N \)
    - Examples: \( S \rightarrow NP \; VP, \; VP \rightarrow VP \; CC \; VP \)
    - Also called rewrites, productions, or local trees

- A PCFG adds:
  - A top-down production probability per rule \( P(Y_1 \; Y_2 \ldots \; Y_k \mid X) \)
( S (NP-SBJ The move)
  (VP followed
    (NP (NP a round)
      (PP of
        (NP (NP similar increases)
          (PP by
            (NP other lenders))
          (PP against
            (NP Arizona real estate loans))))))

,

(S-ADV (NP-SBJ *)
  (VP reflecting
    (NP (NP a continuing decline)
      (PP-LOC in
        (NP that market))))))
Treebank Grammars

- Need a PCFG for broad coverage parsing.
- Can take a grammar right off the trees (doesn’t work well):

```
ROOT → S 1
S → NP VP . 1
NP → PRP 1
VP → VBD ADJP 1
...
```

- Better results by enriching the grammar (e.g., lexicalization).
- Can also get state-of-the-art parsers without lexicalization.
Treebank Grammar Scale

- Treebank grammars can be enormous
  - As FSAs, the raw grammar has ~10K states, excluding the lexicon
  - Better parsers usually make the grammars larger, not smaller
Chomsky normal form:
- All rules of the form $X \rightarrow Y \, Z$ or $X \rightarrow w$
- In principle, this is no limitation on the space of (P)CFGs
  - N-ary rules introduce new non-terminals

- Unaries / empties are “promoted”
- In practice it’s kind of a pain:
  - Reconstructing n-aries is easy
  - Reconstructing unaries is trickier
  - The straightforward transformations don’t preserve tree scores
- Makes parsing algorithms simpler!
CKY Parsing
A Recursive Parser

\[
\text{bestScore}(X,i,j,s) \\
\quad \text{if } (j = i+1) \\
\quad \quad \text{return } \text{tagScore}(X,s[i]) \\
\quad \text{else} \\
\quad \quad \text{return max } \text{score}(X->YZ) \times \\
\quad \quad \quad \text{bestScore}(Y,i,k) \times \\
\quad \quad \quad \text{bestScore}(Z,k,j)
\]

- Will this parser work?
- Why or why not?
- Memory requirements?
A Memoized Parser

- One small change:

```java
bestScore(X,i,j,s)
    if (scores[X][i][j] == null)
        if (j = i+1)
            score = tagScore(X,s[i])
        else
            score = max score(X->YZ) *
                    bestScore(Y,i,k) *
                    bestScore(Z,k,j)
    scores[X][i][j] = score
    return scores[X][i][j]
```
Can also organize things bottom-up

```plaintext
bestScore(s)
  for (i : [0, n-1])
    for (X : tags[s[i]])
      score[X][i][i+1] = tagScore(X, s[i])
    for (diff : [2, n])
      for (i : [0, n-diff])
        j = i + diff
        for (X->YZ : rule)
          for (k : [i+1, j-1])
            score[X][i][j] = max score[X][i][j],
            score(X->YZ) * score[Y][i][k] * score[Z][k][j]
```
Unary Rules

- Unary rules?

```plaintext
bestScore(X,i,j,s)
  if (j = i+1)
    return tagScore(X,s[i])
  else
    return max
      max score(X->YZ) * 
      bestScore(Y,i,k) * 
      bestScore(Z,k,j)
      max score(X->Y) * 
      bestScore(Y,i,j)
```
We need unaries to be non-cyclic

- Can address by pre-calculating the *unary closure*
- Rather than having zero or more unaries, always have exactly one

Alternate unary and binary layers
- Reconstruct unary chains afterwards
bestScoreB(X,i,j,s)  
return max max score(X->YZ) *  
  bestScoreU(Y,i,k) *  
  bestScoreU(Z,k,j)

bestScoreU(X,i,j,s)  
if (j = i+1)  
  return tagScore(X,s[i])  
else  
  return max max score(X->Y) *  
    bestScoreB(Y,i,j)
Analysis
Memory

- **How much memory does this require?**
  - Have to store the score cache
  - Cache size: $|\text{symbols}|*n^2$ doubles
  - For the plain treebank grammar:
    - $X \sim 20K$, $n = 40$, double $\sim 8$ bytes $= \sim 256\text{MB}$
    - Big, but workable.

- **Pruning: Beams**
  - score[$X$][$i$][$j$] can get too large (when?)
  - Can keep beams (truncated maps score[$i$][$j$]) which only store the best few scores for the span [$i,j$]

- **Pruning: Coarse-to-Fine**
  - Use a smaller grammar to rule out most $X[i,j]$
  - Much more on this later…
How much time will it take to parse?

- For each diff (<= n)
  - For each i (<= n)
    - For each rule X → Y Z
      - For each split point k
        Do constant work

- Total time: |rules| * n^3
- Something like 5 sec for an unoptimized parse of a 20-word sentence
- Parsing with the vanilla treebank grammar:

  - Why’s it worse in practice?
    - Longer sentences “unlock” more of the grammar
    - All kinds of systems issues don’t scale

~ 20K Rules
(not an optimized parser!)

Observed exponent: 3.6
Same-Span Reachability

- TOP
- NX
- SQ
- X
- RRC
- ADJP
- ADVP
- FRAG
- INTJ
- NP
- PP
- PRN
- QP
- S
- SBAR
- UCP
- VP
- WHNP
- SINV
- WHADJP
- WHADVP
- SBARQ
- WHPP
- LST
- CONJP
- NAC
Many states are more likely to match larger spans!
Efficient CKY

- **Lots of tricks to make CKY efficient**
  - Some of them are little engineering details:
    - E.g., first choose k, then enumerate through the Y:[i,k] which are non-zero, then loop through rules by left child.
    - Optimal layout of the dynamic program depends on grammar, input, even system details.
  - Another kind is more important (and interesting):
    - Many X[i,j] can be suppressed on the basis of the input string
    - We’ll see this next class as figures-of-merit, A* heuristics, coarse-to-fine, etc
Agenda-Based Parsing
Agenda-Based Parsing

- Agenda-based parsing is like graph search (but over a hypergraph)
- Concepts:
  - Numbering: we number fenceposts between words
  - “Edges” or items: spans with labels, e.g. PP[3,5], represent the sets of trees over those words rooted at that label (cf. search states)
  - A chart: records edges we’ve expanded (cf. closed set)
  - An agenda: a queue which holds edges (cf. a fringe or open set)
Word Items

- Building an item for the first time is called discovery. Items go into the agenda on discovery.
- To initialize, we discover all word items (with score 1.0).

AGENDA

critics[0,1], write[1,2], reviews[2,3], with[3,4], computers[4,5]

CHART [EMPTY]
When we pop a word item, the lexicon tells us the tag item successors (and scores) which go on the agenda.

critics[0,1] write[1,2] reviews[2,3] with[3,4] computers[4,5]
Item Successors

- When we pop items off of the agenda:
  - Graph successors: unary projections (NNS → critics, NP → NNS)
    
    $$Y[i,j] \text{ with } X \rightarrow Y \text{ forms } X[i,j]$$
  
  - Hypergraph successors: combine with items already in our chart
    
    $$Y[i,j] \text{ and } Z[j,k] \text{ with } X \rightarrow Y Z \text{ form } X[i,k]$$

- Enqueue / promote resulting items (if not in chart already)
- Record backtraces as appropriate
- Stick the popped edge in the chart (closed set)

- Queries a chart must support:
  - Is edge $X[i,j]$ in the chart? (What score?)
  - What edges with label $Y$ end at position $j$?
  - What edges with label $Z$ start at position $i$?
An Example

critics write reviews with computers

Sometimes we want to posit nodes in a parse tree that don’t contain any pronounced words:

I want you to parse this sentence
I want [ ] to parse this sentence

These are easy to add to a agenda-based parser!

- For each position $i$, add the “word” edge $\varepsilon[i,i]$
- Add rules like $NP \rightarrow \varepsilon$ to the grammar
- That’s it!
UCS / A*

- With weighted edges, order matters
  - Must expand optimal parse from bottom up (subparses first)
  - CKY does this by processing smaller spans before larger ones
  - UCS pops items off the agenda in order of decreasing Viterbi score
  - A* search also well defined

- You can also speed up the search without sacrificing optimality
  - Can select which items to process first
  - Can do with any “figure of merit” [Charniak 98]
  - If your figure-of-merit is a valid A* heuristic, no loss of optimality [Klein and Manning 03]
There was nothing magical about words spanning exactly one position.

When working with speech, we generally don’t know how many words there are, or where they break.

We can represent the possibilities as a lattice and parse these just as easily.
Learning PCFGs
Treebank PCFGs

- Use PCFGs for broad coverage parsing
- Can take a grammar right off the trees (doesn’t work well):

```
ROOT → S
S → NP VP .
NP → PRP
VP → VBD ADJP
```

Model | F1  
------|-----
Baseline | 72.0
Not every NP expansion can fill every NP slot

- A grammar with symbols like “NP” won’t be context-free
- Statistically, conditional independence too strong
Non-Independence

- Independence assumptions are often too strong.

- Example: the expansion of an NP is highly dependent on the parent of the NP (i.e., subjects vs. objects).
- Also: the subject and object expansions are correlated!
Grammar Refinement

- **Example:** PP attachment

```
They

VP

raised

a point of order

NP
```
Grammar Refinement

- Structure Annotation [Johnson ’98, Klein&Manning ’03]
- Lexicalization [Collins ’99, Charniak ’00]
- Latent Variables [Matsuzaki et al. 05, Petrov et al. ’06]

```
S
   NP^she
       PRP VBD NP^noise
          She heard DT NN
             the noise
```
Structural Annotation
The Game of Designing a Grammar

- Annotation refines base treebank symbols to improve statistical fit of the grammar
  - Structural annotation
Typical Experimental Setup

- **Corpus:** Penn Treebank, WSJ

- **Accuracy – F1:** harmonic mean of per-node labeled precision and recall.

- **Here:** also size – number of symbols in grammar.

| Training: | sections | 02-21 |
| Development: | section | 22 (here, first 20 files) |
| Test: | section | 23 |
Vertical Markov Order:
- Vertical Markov order: rewrites depend on past $k$ ancestor nodes. (cf. parent annotation)

Order 1
- S
  - NP
    - PRP
      - He
    - VBD
      - was
    - ADJP
      - right
  - VP

Order 2
- S
  - NP
    - PRP
      - He
    - VBD
      - was
  - VP
    - ADJP
      - right

Bar charts:
- Vertical Markov Order:
  - 72%: 1
  - 73%: 2v
  - 74%: 2
  - 75%: 3v
  - 76%: 3
- Symbols:
  - 72%: 1
  - 73%: 2v
  - 74%: 2
  - 75%: 3v
  - 76%: 3

Graphs:
- Vertical Markov Order
- Symbols
Horizontal Markovization

Order 1

Order ∞

Symbols

Horizontal Markov Order

Horizontal Markov Order
Unary Splits

- Problem: unary rewrites used to transmute categories so a high-probability rule can be used.

- Solution: Mark unary rewrite sites with -U

<table>
<thead>
<tr>
<th>Annotation</th>
<th>F1</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>77.8</td>
<td>7.5K</td>
</tr>
<tr>
<td>UNARY</td>
<td>78.3</td>
<td>8.0K</td>
</tr>
</tbody>
</table>
Problem: Treebank tags are too coarse.

Example: Sentential, PP, and other prepositions are all marked IN.

Partial Solution:
- Subdivide the IN tag.

<table>
<thead>
<tr>
<th>Annotation</th>
<th>F1</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous</td>
<td>78.3</td>
<td>8.0K</td>
</tr>
<tr>
<td>SPLIT-IN</td>
<td>80.3</td>
<td>8.1K</td>
</tr>
</tbody>
</table>
A Fully Annotated (Unlex) Tree

```
ROOT
  \|-- S^ROOT-v
         \|-- "^SNP^S-B
                   \|-- "DT-U^NP
                           | This
                   \|-- VBZ^BE^VP
                           | is
                   \|-- NP^VP-B
                           \|-- NN^NP
                                   panic
                           \|-- NN^NP
                                   buying
```
Some Test Set Results

<table>
<thead>
<tr>
<th>Parser</th>
<th>LP</th>
<th>LR</th>
<th>F1</th>
<th>CB</th>
<th>0 CB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magerman 95</td>
<td>84.9</td>
<td>84.6</td>
<td><strong>84.7</strong></td>
<td>1.26</td>
<td>56.6</td>
</tr>
<tr>
<td>Collins 96</td>
<td>86.3</td>
<td>85.8</td>
<td><strong>86.0</strong></td>
<td>1.14</td>
<td>59.9</td>
</tr>
<tr>
<td>Unlexicalized</td>
<td>86.9</td>
<td>85.7</td>
<td><strong>86.3</strong></td>
<td>1.10</td>
<td>60.3</td>
</tr>
<tr>
<td>Charniak 97</td>
<td>87.4</td>
<td>87.5</td>
<td><strong>87.4</strong></td>
<td>1.00</td>
<td>62.1</td>
</tr>
<tr>
<td>Collins 99</td>
<td>88.7</td>
<td>88.6</td>
<td><strong>88.6</strong></td>
<td>0.90</td>
<td>67.1</td>
</tr>
</tbody>
</table>

- Beats “first generation” lexicalized parsers.
- Lots of room to improve – more complex models next.
Efficient Parsing for Structural Annotation
Grammar Projections

Coarse Grammar

```
S
   NP   VP
      PRP  VBD  ADJP
         He  was  right
```

```
NP \rightarrow DT N'
```

Fine Grammar

```
S'\text{ROOT}
   NP'S   VP'S
      PRP  VBD  ADVP\text{VP}
         He  was  right
```

```
NP'S \rightarrow DT^NP N'\ldots DT^NP
```

Note: X-Bar Grammars are projections with rules like $XP \rightarrow Y X'$ or $XP \rightarrow X' Y$ or $X' \rightarrow X$
Coarse-to-Fine Pruning

For each coarse chart item $X[i,j]$, compute posterior probability:

$$\frac{P_{IN}(X, i, j) \cdot P_{OUT}(X, i, j)}{P_{IN}(root, 0, n)} < \text{threshold}$$

E.g. consider the span 5 to 12:
Computing (Max-)Marginals

\[ \beta(x_{ij}, y) = \sum_{x} \beta(x_{ij}, x) \cdot \beta(x, y) \]

\[ \beta(x_{ij}, x_{ik}) \cdot \beta(x_{ij}, x_{jk}) \]
Inside and Outside Scores

\[ \alpha(x, i, j) = \sum_{A \to B} \beta(B, k, j) \]

\[ \alpha(A, k, j) \]

\[ \beta(B, k, i) \]

\[ c \to z \]

Source: [Image of handwritten notes](image)
Pruning with A*

- You can also speed up the search without sacrificing optimality
- For agenda-based parsers:
  - Can select which items to process first
  - Can do with any “figure of merit” [Charniak 98]
  - If your figure-of-merit is a valid A* heuristic, no loss of optimiality [Klein and Manning 03]
A* Parsing

<table>
<thead>
<tr>
<th>Estimate</th>
<th>SX</th>
<th>SXL</th>
<th>SXLRF</th>
<th>TRUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary</td>
<td>(1,6,NP)</td>
<td>(1,6,NP,VBZ)</td>
<td>(1,6,NP,VBZ,“,”)</td>
<td>(entire context)</td>
</tr>
<tr>
<td>Best Tree</td>
<td><img src="image1" alt="Tree Diagram" /></td>
<td><img src="image2" alt="Tree Diagram" /></td>
<td><img src="image3" alt="Tree Diagram" /></td>
<td><img src="image4" alt="Tree Diagram" /></td>
</tr>
<tr>
<td>Score</td>
<td>-11.3</td>
<td>-13.9</td>
<td>-15.1</td>
<td>-18.1</td>
</tr>
</tbody>
</table>
Lexicalization
The Game of Designing a Grammar

- Annotation refines base treebank symbols to improve statistical fit of the grammar
  - Structural annotation [Johnson ’98, Klein and Manning 03]
  - Head lexicalization [Collins ’99, Charniak ’00]
Problems with PCFGs

- If we do no annotation, these trees differ only in one rule:
  - VP → VP PP
  - NP → NP PP
- Parse will go one way or the other, regardless of words
- We addressed this in one way with unlexicalized grammars (how?)
- Lexicalization allows us to be sensitive to specific words
Problems with PCFGs

- What’s different between basic PCFG scores here?
- What (lexical) correlations need to be scored?
Lexicalized Trees

- Add “head words” to each phrasal node
  - Syntactic vs. semantic heads
  - Headship not in (most) treebanks
  - Usually use head rules, e.g.:
    - **NP:**
      - Take leftmost NP
      - Take rightmost N*
      - Take rightmost JJ
      - Take right child
    - **VP:**
      - Take leftmost VB*
      - Take leftmost VP
      - Take left child
Lexicalized PCFGs?

- Problem: we now have to estimate probabilities like

  \[ \text{VP(saw)} \rightarrow \text{VBD(saw)} \text{ NP-C(her) NP(today)} \]

- Never going to get these atomically off of a treebank

- Solution: break up derivation into smaller steps
Lexical Derivation Steps

- **A derivation of a local tree [Collins 99]**

  
  \[ \text{VP(\text{saw})} \]

  \[ \rightarrow \]

  \[ \text{VBD(\text{saw})} \]

  
  \[ \text{VP(\text{saw})} \]

  \[ \rightarrow \]

  \[ \text{VBD(\text{saw}) \{NP-C(\quad)\}} \]

  
  \[ \text{VP(\text{saw})} \]

  \[ \rightarrow \]

  \[ \text{VBD(\text{saw}) \quad NP-C(\quad) \quad NP(\quad)} \]

  
  \[ \text{VP(\text{saw})} \]

  \[ \rightarrow \]

  \[ \text{VBD(\text{saw}) \quad NP-C(\text{her}) \quad NP(\text{today})} \]

  
  Choose a head tag and word

  Choose a complement bag

  Generate children (incl. adjuncts)

  Recursively derive children
Lexicalized CKY

bestScore(X, i, j, h)
  if (j = i+1)
    return tagScore(X, s[i])
  else
    return
      \[
      \max_{k,h',X \rightarrow YZ} \ \max \ \text{score}(X[h] \rightarrow Y[h] \ Z[h']) \ * \\
      \text{bestScore}(Y, i, k, h) \ * \\
      \text{bestScore}(Z, k, j, h')
      \]
      \[
      \max_{k,h',X \rightarrow YZ} \ \text{score}(X[h] \rightarrow Y[h'] \ Z[h]) \ * \\
      \text{bestScore}(Y, i, k, h') \ * \\
      \text{bestScore}(Z, k, j, h)
      \]
Efficient Parsing for Lexical Grammars
Turns out, you can do (a little) better [Eisner 99]

- Gives an $O(n^4)$ algorithm
- Still prohibitive in practice if not pruned
Pruning with Beams

- The Collins parser prunes with per-cell beams [Collins 99]
  - Essentially, run the $O(n^5)$ CKY
  - Remember only a few hypotheses for each span $<i,j>$. 
  - If we keep $K$ hypotheses at each span, then we do at most $O(nK^2)$ work per span (why?)
  - Keeps things more or less cubic (and in practice is more like linear!)

- Also: certain spans are forbidden entirely on the basis of punctuation (crucial for speed)
The Charniak parser prunes using a two-pass, coarse-to-fine approach [Charniak 97+]

- First, parse with the base grammar
- For each X:[i,j] calculate $P(X|i,j,s)$
  - This isn’t trivial, and there are clever speed ups
- Second, do the full $O(n^5)$ CKY
  - Skip any X :[i,j] which had low (say, < 0.0001) posterior
  - Avoids almost all work in the second phase!

- Charniak et al 06: can use more passes
- Petrov et al 07: can use many more passes
Results

- **Some results**
  - Collins 99 – 88.6 F1 (generative lexical)
  - Charniak and Johnson 05 – 89.7 / 91.3 F1 (generative lexical / reranked)
  - Petrov et al 06 – 90.7 F1 (generative unlexical)
  - McClosky et al 06 – 92.1 F1 (gen + rerank + self-train)

- **However**
  - Bilexical counts rarely make a difference (why?)
  - Gildea 01 – Removing bilexical counts costs < 0.5 F1
Latent Variable PCFGs
The Game of Designing a Grammar

- Annotation refines base treebank symbols to improve statistical fit of the grammar
  - Parent annotation [Johnson ’98]
  - Head lexicalization [Collins ’99, Charniak ’00]
  - Automatic clustering?
Latent Variable Grammars

Parse Tree

Sentence

Derivations $t : T$

Parameters $\theta$
Learning Latent Annotations

EM algorithm:
- Brackets are known
- Base categories are known
- Only induce subcategories

Just like Forward-Backward for HMMs.
Refinement of the DT tag

DT

- the (0.50)
- a (0.24)
- The (0.08)

a (0.61)
the (0.19)
an (0.11)

the (0.80)
The (0.15)
a (0.01)

this (0.39)
that (0.28)
That (0.11)

some (0.20)
all (0.19)
those (0.12)

DT-1 | DT-2 | DT-3 | DT-4
Hierarchical refinement

- the (0.50)
  - a (0.24)
  - the (0.08)
    - the (0.54)
      - a (0.25)
      - The (0.09)
      - a (0.61)
        - the (0.19)
        - an (0.11)
    - that (0.15)
      - this (0.14)
      - some (0.11)
      - this (0.39)
        - that (0.28)
        - That (0.11)
      - some (0.20)
        - all (0.19)
        - those (0.12)
Hierarchical Estimation Results

<table>
<thead>
<tr>
<th>Model</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Training</td>
<td>87.3</td>
</tr>
<tr>
<td>Hierarchical Training</td>
<td>88.4</td>
</tr>
</tbody>
</table>
Refinement of the , tag

- Splitting all categories equally is wasteful:
Adaptive Splitting

- Want to split complex categories more
- Idea: split everything, roll back splits which were least useful
Adaptive Splitting Results

Model | F1
--- | ---
Previous | 88.4
With 50% Merging | 89.5
Number of Phrasal Subcategories

![Graph showing the number of phrasal subcategories]
Learned Splits

- **Proper Nouns (NNP):**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NNP-12</td>
<td>John</td>
<td>Robert</td>
<td>James</td>
</tr>
<tr>
<td>NNP-2</td>
<td>J.</td>
<td>E.</td>
<td>L.</td>
</tr>
<tr>
<td>NNP-1</td>
<td>Bush</td>
<td>Noriega</td>
<td>Peters</td>
</tr>
<tr>
<td>NNP-15</td>
<td>New</td>
<td>San</td>
<td>Wall</td>
</tr>
<tr>
<td>NNP-3</td>
<td>York</td>
<td>Francisco</td>
<td>Street</td>
</tr>
</tbody>
</table>

- **Personal pronouns (PRP):**

<table>
<thead>
<tr>
<th>PRP-0</th>
<th>it</th>
<th>He</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRP-1</td>
<td>it</td>
<td>he</td>
<td>they</td>
</tr>
<tr>
<td>PRP-2</td>
<td>it</td>
<td>them</td>
<td>him</td>
</tr>
</tbody>
</table>
Learned Splits

- Relative adverbs (RBR):

<table>
<thead>
<tr>
<th>RBR</th>
<th>further</th>
<th>lower</th>
<th>higher</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBR-0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBR-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBR-2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Cardinal Numbers (CD):

<table>
<thead>
<tr>
<th>CD</th>
<th>one</th>
<th>two</th>
<th>Three</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD-7</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>CD-4</td>
<td>1989</td>
<td>1990</td>
<td>1988</td>
</tr>
<tr>
<td>CD-11</td>
<td>billion</td>
<td></td>
<td>trillion</td>
</tr>
<tr>
<td>CD-0</td>
<td>1</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>CD-3</td>
<td>1</td>
<td>30</td>
<td>31</td>
</tr>
<tr>
<td>CD-9</td>
<td>78</td>
<td>58</td>
<td>34</td>
</tr>
</tbody>
</table>
Final Results (Accuracy)

<table>
<thead>
<tr>
<th>Language</th>
<th>Method</th>
<th>≤ 40 words F1</th>
<th>all F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENG</td>
<td>Charniak&amp;Johnson ‘05 (generative)</td>
<td>90.1</td>
<td>89.6</td>
</tr>
<tr>
<td></td>
<td>Split / Merge</td>
<td>90.6</td>
<td>90.1</td>
</tr>
<tr>
<td>GER</td>
<td>Dubey ‘05</td>
<td>76.3</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Split / Merge</td>
<td>80.8</td>
<td>80.1</td>
</tr>
<tr>
<td>CHN</td>
<td>Chiang et al. ‘02</td>
<td>80.0</td>
<td>76.6</td>
</tr>
<tr>
<td></td>
<td>Split / Merge</td>
<td>86.3</td>
<td>83.4</td>
</tr>
</tbody>
</table>

Still higher numbers from reranking / self-training methods
Efficient Parsing for Hierarchical Grammars
Coarse-to-Fine Inference

- Example: PP attachment

```
S
  NP
  |
  PRP
  |
  They

S
  NP
  |
  VP
  |
  ???

V
  raised
  |
  DT
  a
  |
  NN
  point

IN
  of
  |
  NP
  order
```
Hierarchical Pruning

coarse: ... QP NP VP ...

split in two: ... QP1 QP2 NP1 NP2 VP1 VP2 ...

split in four: ... QP1 QP1 QP3 QP4 NP1 NP2 NP3 NP4 VP1 VP2 VP3 VP4 ...

split in eight: ... ...

...
Bracket Posteriors
1621 min
111 min
35 min
15 min
(no search error)
Unsupervised Tagging
Unsupervised Tagging?

- **AKA part-of-speech induction**

- **Task:**
  - Raw sentences in
  - Tagged sentences out

- **Obvious thing to do:**
  - Start with a (mostly) uniform HMM
  - Run EM
  - Inspect results
EM for HMMs: Process

- Alternate between recomputing distributions over hidden variables (the tags) and reestimating parameters
- Crucial step: we want to tally up how many (fractional) counts of each kind of transition and emission we have under current params:

\[
\text{count}(w, s) = \sum_{i: w_i = w} P(t_i = s | w)
\]

\[
\text{count}(s \to s') = \sum_i P(t_{i-1} = s, t_i = s' | w)
\]

- Same quantities we needed to train a CRF!
Some (discouraging) experiments [Merialdo 94]

Setup:
- You know the set of allowable tags for each word
- Fix k training examples to their true labels
  - Learn $P(w|t)$ on these examples
  - Learn $P(t|t_{-1}, t_{-2})$ on these examples
- On n examples, re-estimate with EM

Note: we know allowed tags but not frequencies
### Merialdo: Results

<table>
<thead>
<tr>
<th>Iter</th>
<th>Correct tags (% words) after ML on 1M words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>