NLP
Introduction to NLP

Unlexicalized Parsing
• One extreme
  – PTB categories

• Another extreme
  – individual words

• Try something in the middle
  – e.g., collapse table/N and house/N
  – but don’t collapse of/PP and in/PP
Markovization

- Markovization
  - horizontal
  - vertical

![Figure 2: Markovizations: F₁ and grammar size.](image)
Tag Splitting

- Split some part-of-speech tags into tags marked with semantic function labels
- E.g., IN–SBAR

Figure 5: An error resolved with the TAG-PA annotation (of the IN tag): (a) the incorrect baseline parse and (b) the correct TAG-PA parse. SPLIT-IN also resolves this error.
State Splitting

- Example:
  - Distinguish between “and” and “but” from the other CCs
State Splitting

- **Example:**
  - Distinguish between “and” and “but” from the other CCs

- **Done manually**
  - So hard to port

Figure 7: An error resolved with the SPLIT-VP annotation: (a) the incorrect baseline parse and (b) the correct SPLIT-VP parse.
Petrov et al. (2006)

• Replace each nonterminal A by a set of categories $A[X_1]...A[X_n]$
• $X_1..X_n$ are latent variables
• EM–based
• Iterative splits and merges
  – Split as long as accuracy improves
  – Merge if accuracy doesn’t improve
NLP