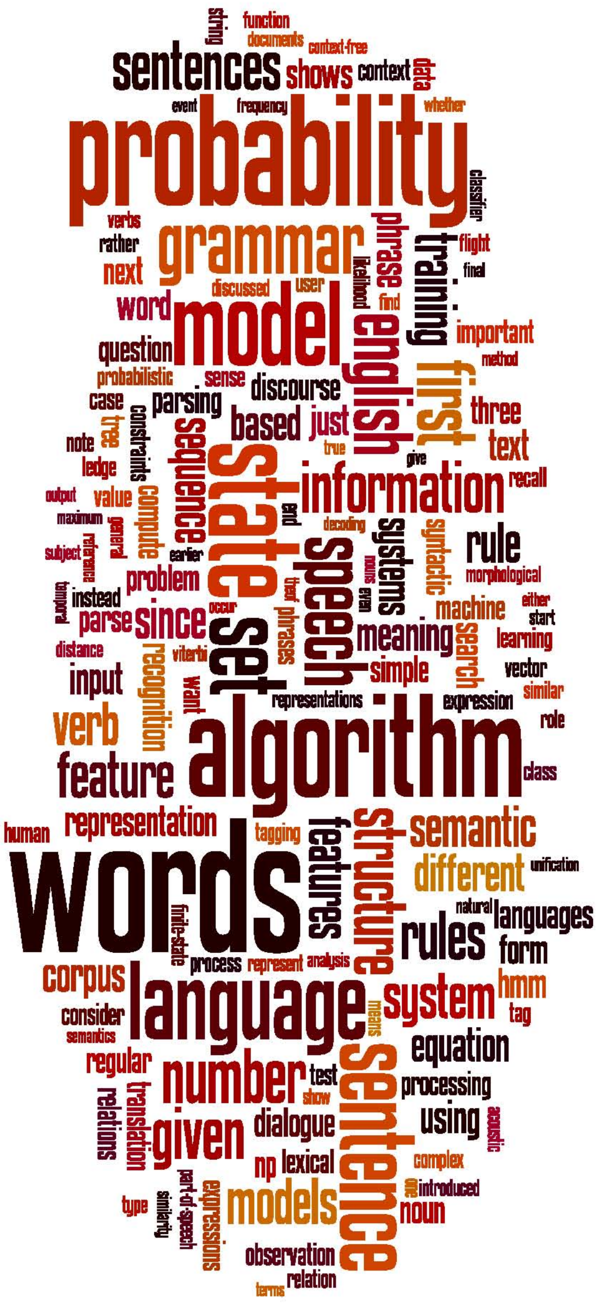


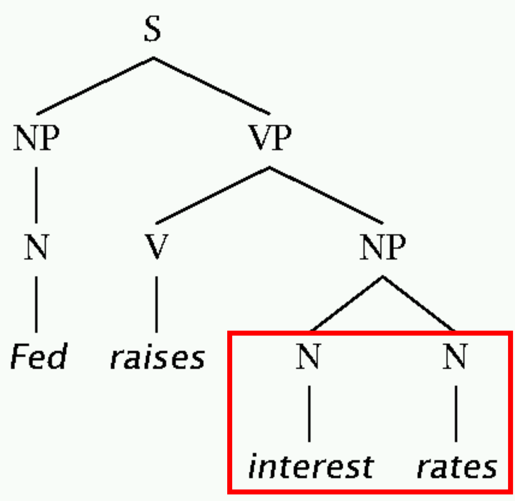
Statistical Natural Language Parsing

Two views of syntactic structure

Two views of linguistic structure: 1. Constituency (phrase structure)

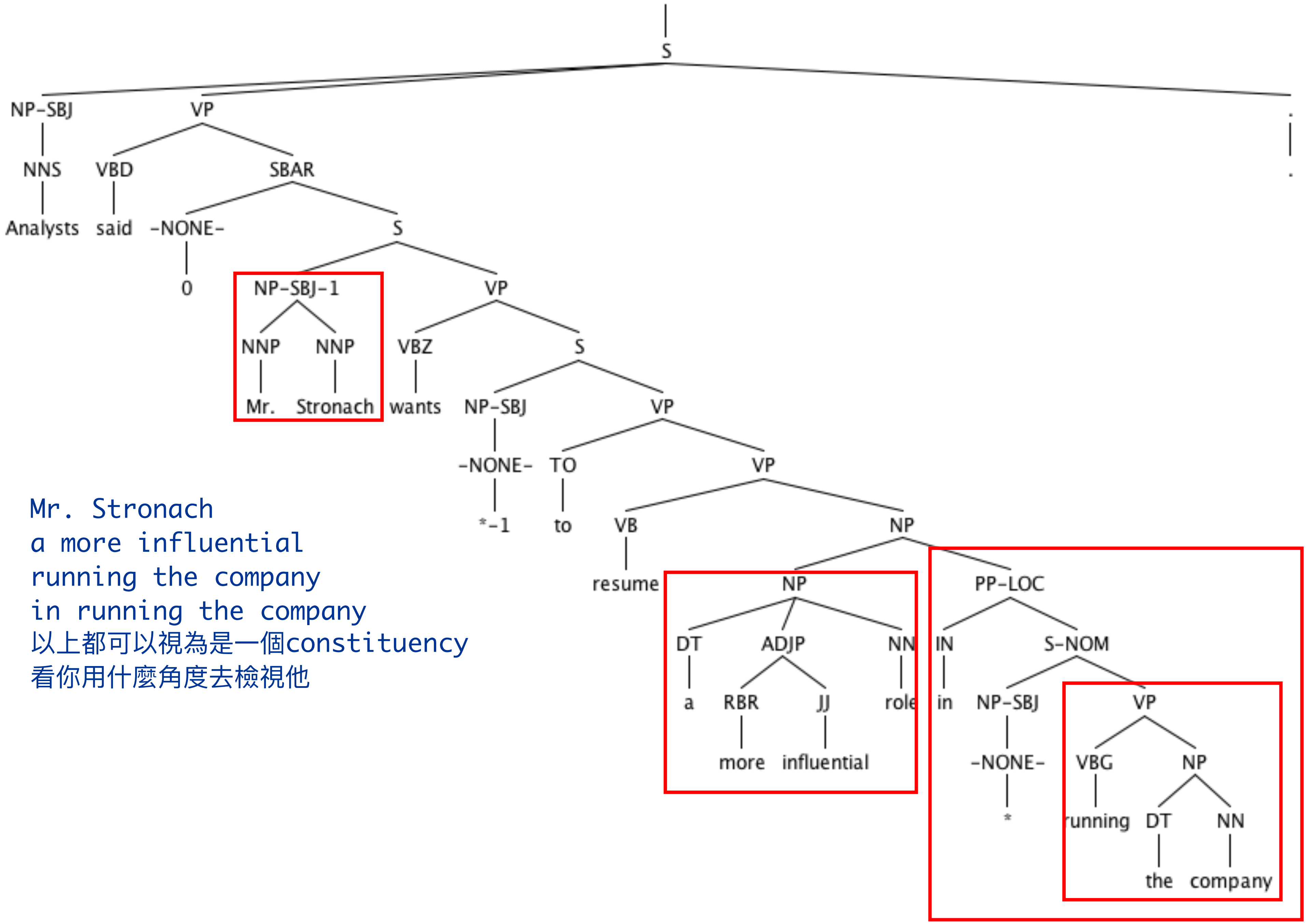
除了標記一句話的磁性之外，還要標記出子句的組合

- Phrase structure organizes words into nested constituents.
- How do we know what is a constituent? (Not that linguists don't argue about some cases.) 成分
  - Distribution: a constituent behaves as a unit that can appear in different places: to the children, about drugs是兩個units 在句子中換位置是可以的，但拆開就不行
    - John talked [to the children] [about drugs].
    - John talked [about drugs] [to the children].
    - *John talked drugs to the children about
  - Substitution/expansion/pro-forms:
    - I sat [on the box/right on top of the box/there].
  - Coordination, regular internal structure, no intrusion, fragments, semantics, …



Two views of linguistic structure: 1. Constituency (phrase structure)

- Phrase structure organizes words into nested constituents.
- How do we know what is a constituent? (Not that linguists don’t argue about some cases.)
    - Distribution: a constituent behaves as a unit that can appear in different places:
        - John talked [to the children] [about drugs].
        - John talked [about drugs] [to the children].
        - *John talked drugs to the children about
    - Substitution/expansion/pro-forms:
        - I sat [on the box/right on top of the box/there].
    - Coordination, regular internal structure, no intrusion, fragments, semantics, …

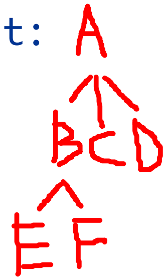


Mr. Stronach
a more influential
running the company
in running the company
以上都可以視為是一個constituency
看你用什麼角度去檢視他

Headed phrase structure

- VP → … VB* …
- NP → … NN* …
- ADJP → … JJ* …
- ADVP → … RB* …

```
e.g.
A -> BCD
B -> EF
input : A
output:
```



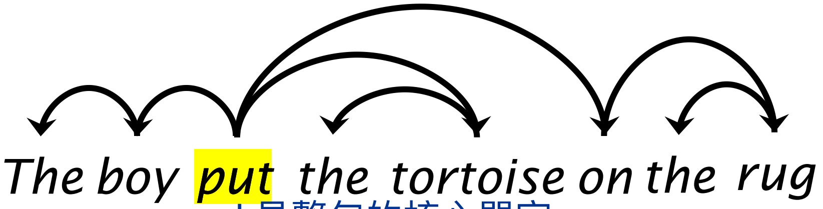
- SBAR(Q) → S|SINV|SQ → … NP VP …

- Plus minor phrase types:
  - QP (quantifier phrase in NP), CONJP (multi word constructions: *as well as*), INTJ (interjections), etc.

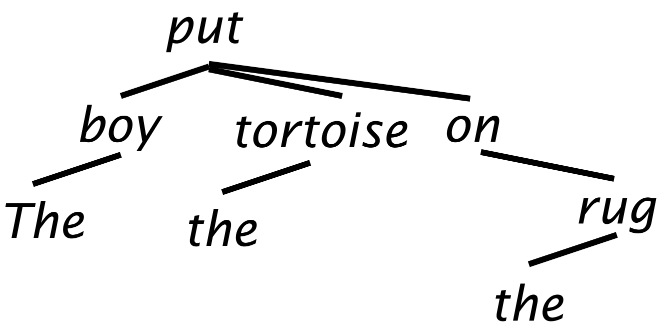
Two views of linguistic structure: 2. Dependency structure

- Dependency structure shows which words depend on (modify or are arguments of) which other words.

不是用整個句子的詞性文法結構來看
而是看字與字之間的關係
建立出一個Context Free Tree



The boy put the tortoise on the rug



put
boy tortoise on
The the rug
the

put是整句的核心單字
put有三個參數: put(boy, tortoise, rug)
boy(the)
tortoise(the)
rug(the)
on(rug)

Two views of linguistic structure: 2. Dependency structure

- Dependency structure shows which words depend on (modify or are arguments of) which other words.

The boy put the tortoise on the rug

Statistical Natural Language Parsing

Two views of syntactic structure

Statistical Natural Language Parsing

Parsing: The rise of data and statistics

Pre 1990 (“Classical”) NLP Parsing

1990前，手寫文法規則，直接找出句子的結構
但效果很差，原因是文法結構太多變，根本寫不完
再來是一般的文章的文法也不一定完全正確

- Wrote symbolic grammar (CFG or often richer) and lexicon

  S → NP VP
  NP → (DT) NN
  NP → NN NNS
  NP → NNP
  VP → V NP

  NN → *interest*
  NNS → *rates*
  NNS → *raises*
  VBP → *interest*
  VBZ → *rates*

- Used grammar/proof systems to prove parses from words
- This scaled very badly and didn’t give coverage. For sentence:

  Fed raises interest rates 0.5% in effort to control inflation

  - Minimal grammar: 36 parses
  - Simple 10 rule grammar: 592 parses
  - Real-size broad-coverage grammar: millions of parses

Classical NLP Parsing: The problem and its solution

如果文法規則設計的比較嚴，那會有一些句子無法分析
如果文法規則設計的比較鬆，那一個句子可能會有很多種語法解釋，而系統卻無從挑選

- Categorical constraints can be added to grammars to limit unlikely/weird parses for sentences
  - But the attempt make the grammars not robust
    - In traditional systems, commonly 30% of sentences in even an edited text would have *no* parse.
- A less constrained grammar can parse more sentences
  - But simple sentences end up with ever more parses with no way to choose between them
- We need mechanisms that allow us to find the most likely parse(s) for a sentence
  - Statistical parsing lets us work with very loose grammars that admit millions of parses for sentences but still quickly find the best parse(s)

The rise of annotated data: The Penn Treebank

[Marcus et al. 1993, *Computational Linguistics*]

用大量資料，配上Treebank，可以做有效的parse
有了Treebank，就可以產生每一個句子的Treebank
就可以丟到機器學習Model去訓練

```
( (S
    (NP-SBJ (DT The) (NN move))
    (VP (VBD followed)
      (NP
        (NP (DT a) (NN round))
        (PP (IN of)
          (NP
            (NP (JJ similar) (NNS increases))
            (PP (IN by)
              (NP (JJ other) (NNS lenders)))
            (PP (IN against)
              (NP (NNP Arizona) (JJ real) (NN estate) (NNS loans))))))
      (, ,)
      (S-ADV
        (NP-SBJ (-NONE- *))
        (VP (VBG reflecting)
          (NP
            (NP (DT a) (VBG continuing) (NN decline))
            (PP-LOC (IN in)
              (NP (DT that) (NN market)))))))
    (. .)))
```

The rise of annotated data

- Starting off, building a treebank seems a lot slower and less useful than building a grammar

- But a treebank gives us many things
  - Reusability of the labor
    - Many parsers, POS taggers, etc.
    - Valuable resource for linguistics
  - Broad coverage
  - Frequencies and distributional information
  - A way to evaluate systems

Statistical parsing applications

Statistical parsers are now robust and widely used in larger NLP applications:

- High precision question answering [Pasca and Harabagiu SIGIR 2001]
- Improving biological named entity finding [Finkel et al. JNLPBA 2004]
- Syntactically based sentence compression [Lin and Wilbur 2007]
- Extracting opinions about products [Bloom et al. NAACL 2007]
- Improved interaction in computer games [Gorniak and Roy 2005]
- Helping linguists find data [Resnik et al. BLS 2005]
- Source sentence analysis for machine translation [Xu et al. 2009]
- Relation extraction systems [Fundel et al. *Bioinformatics* 2006]

Statistical Natural Language Parsing

Parsing: The rise of data and statistics

Statistical Natural Language Parsing

An exponential number of attachments

Attachment ambiguities

Attachment = Connection

- A key parsing decision is how we 'attach' various constituents
  - PPs, adverbial or participial phrases, infinitives, coordinations, etc.

The board approved [its acquisition] [by Royal Trustco Ltd.]

[of Toronto]

[for $27 a share]

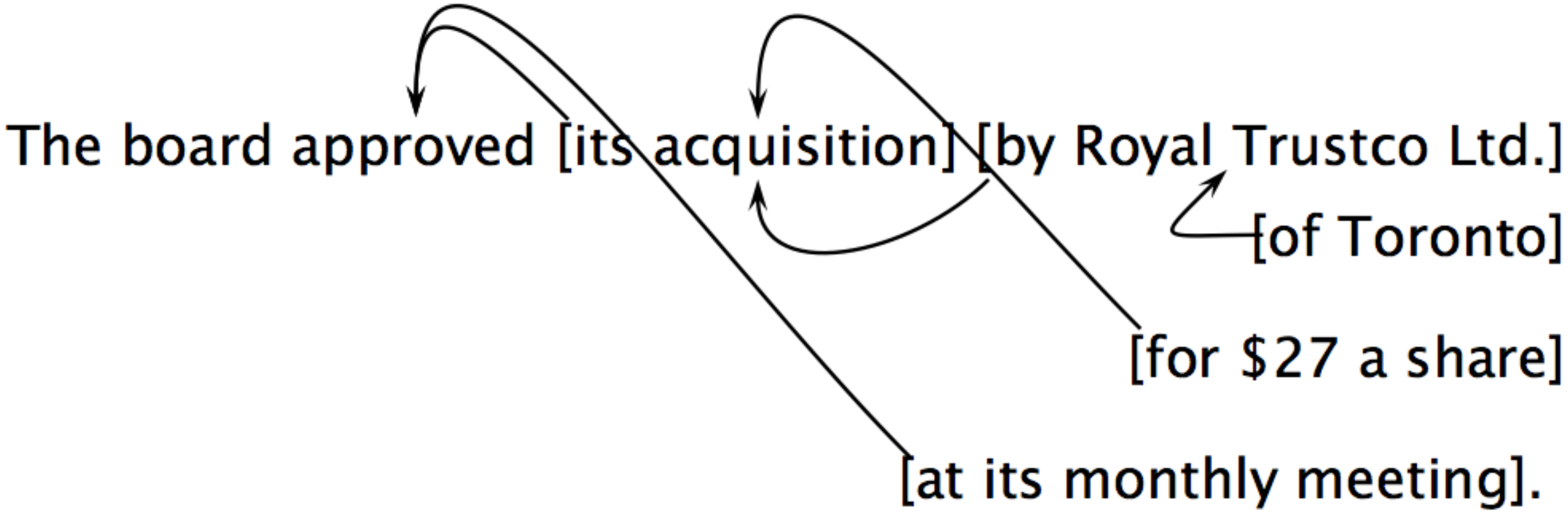
[at its monthly meeting].

假設現在已經切好句子的每一個constituents
那要怎麼知道每一個constituent之間的關聯？
但如果有n個constituents
就會有n^n種可能

- Catalan numbers: $C_n = (2n)!/[(n+1)!n!]$
- An exponentially growing series, which arises in many tree-like contexts:
  - E.g., the number of possible triangulations of a polygon with n+2 sides
    - Turns up in triangulation of probabilistic graphical models….

Attachment ambiguities

- A key parsing decision is how we ‘attach’ various constituents
  - PPs, adverbial or participial phrases, infinitives, coordinations, etc.



The board approved [its acquisition] [by Royal Trustco Ltd.]

[of Toronto]

[for $27 a share]

[at its monthly meeting].

- Catalan numbers: $C_n = (2n)!/[(n+1)!n!]$
- An exponentially growing series, which arises in many tree-like contexts:
  - E.g., the number of possible triangulations of a polygon with $n+2$ sides
    - Turns up in triangulation of probabilistic graphical models….

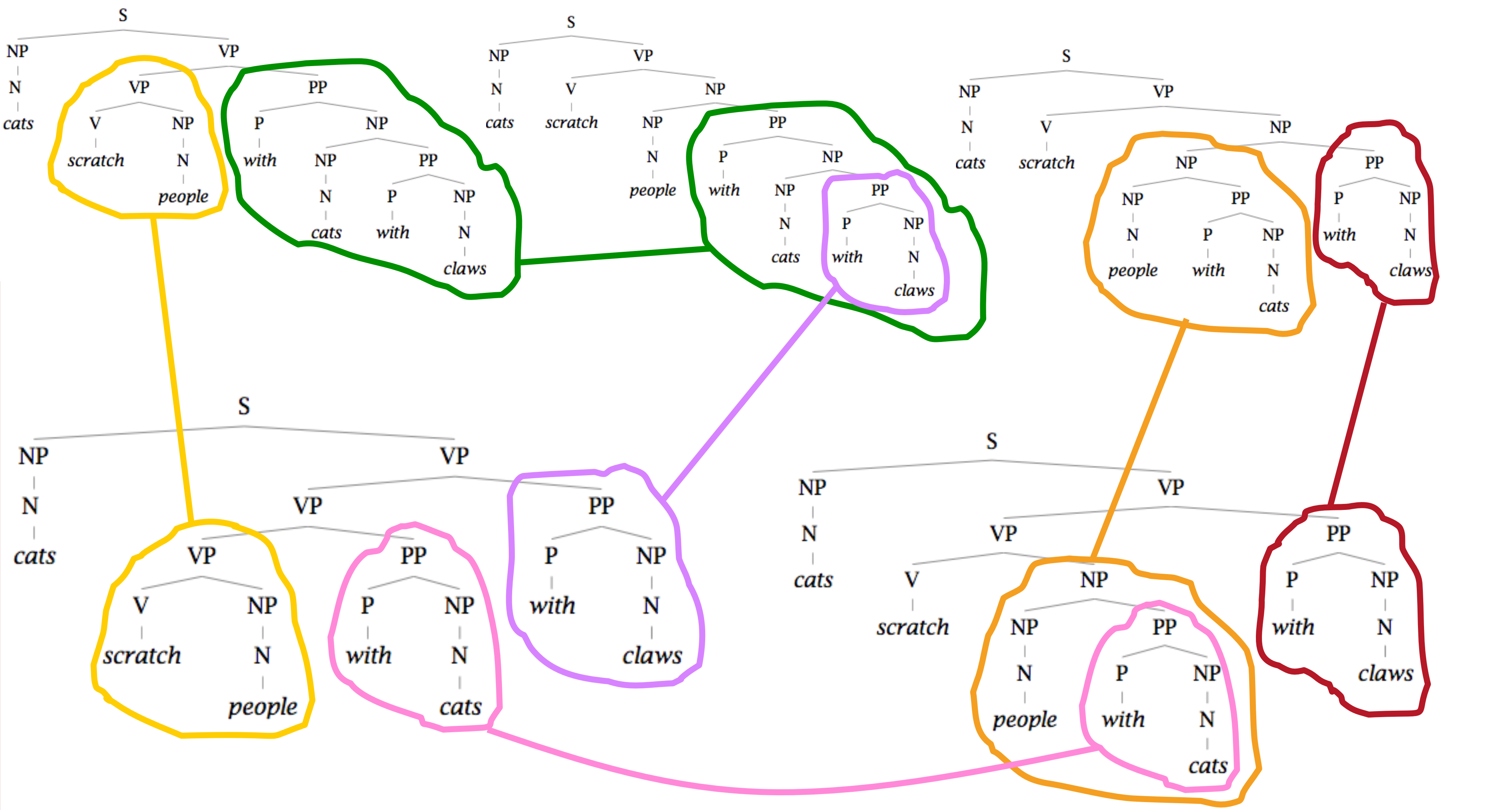
Quiz Question!

- How many distinct parses does the following sentence have due to PP attachment ambiguities?
  - A PP can attach to any preceding V or N within the verb phrase, subject only to the parse still being a tree.
    - (This is equivalent to there being no crossing dependencies, where if d_2 is a dependent of d_1 and d_3 is a dependent of d_2, then the line d_2–d_3 begins at d_2 under the line from d_1 to d_2.)

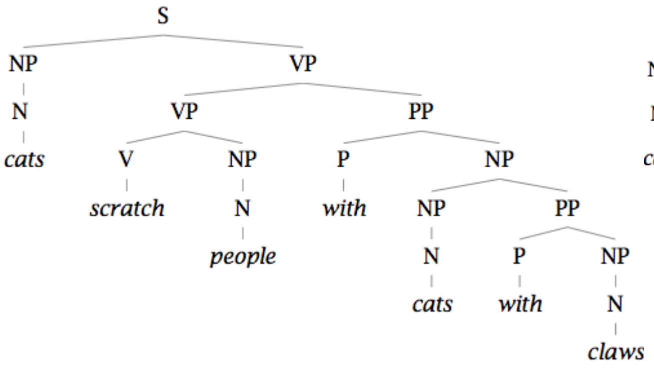
John wrote the book with a pen in the room.

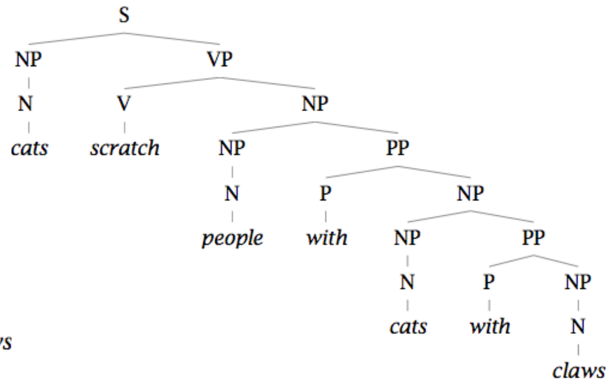
Two problems to solve:
1. Repeated work…

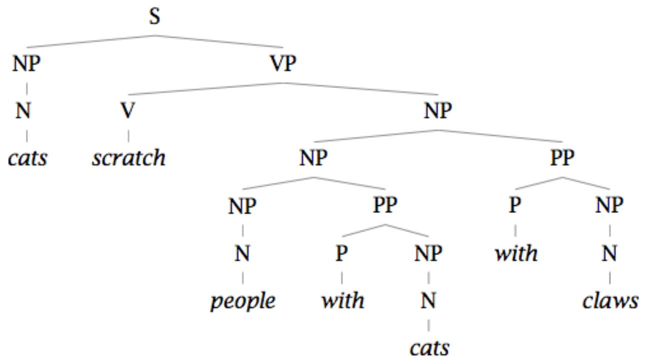
相同顏色圈起來的部分
代表的都是一模一樣的constituents
如果要產生該句子的所有Attachment Ambiguities
就會不斷地重複運算相同的部分
一個好的Parser應該要能避免重複運算
做到類似Dynamic Programming?

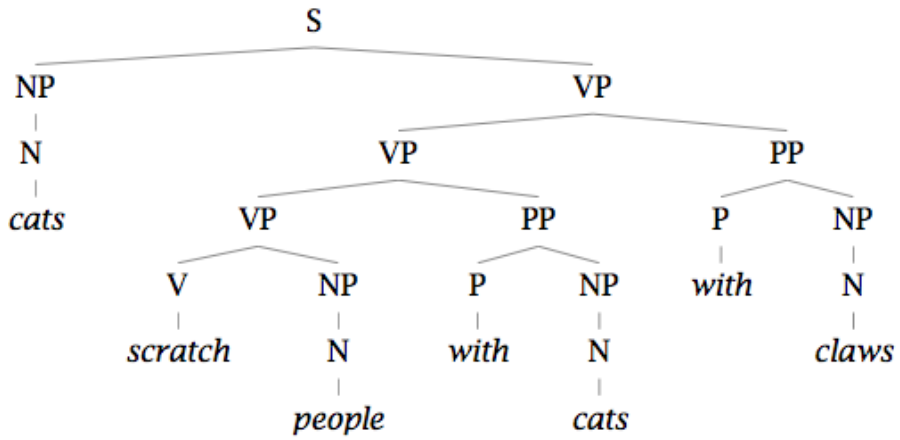


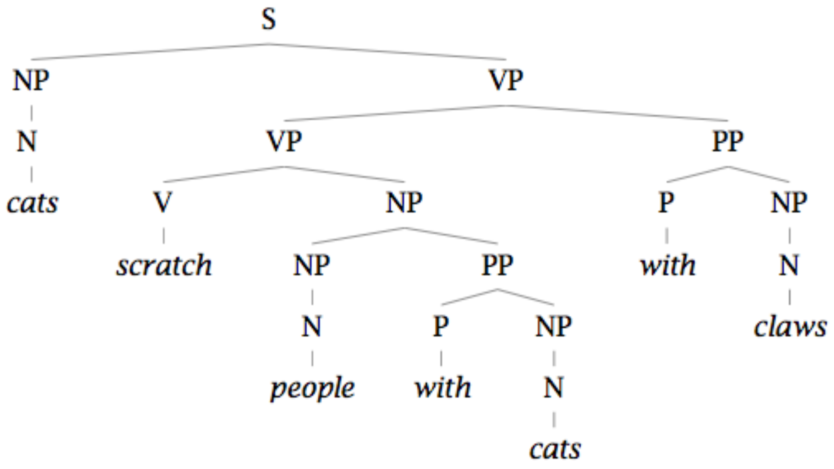
Two problems to solve: 1. Repeated work…











Two problems to solve:
2. Choosing the correct parse

- How do we work out the correct attachment:
  - She saw the man with a telescope
- Is the problem 'AI complete'? Yes, but …
- Words are good predictors of attachment
  - Even absent full understanding
  - Moscow sent more than 100,000 soldiers into Afghanistan …
  - Sydney Water breached an agreement with NSW Health …
- Our statistical parsers will try to exploit such statistics.

Statistical Natural Language Parsing

An exponential number of attachments