統計的手法による
文法モデリングと構文解析

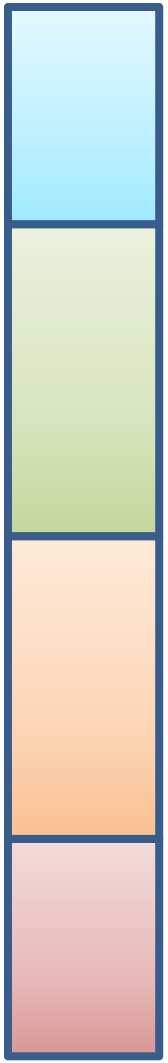
進藤　裕之

NTT コミュニケーション科学基礎研究所

2012.12.19

最先端構文解析とその周辺@統計数理研究所

全体構成



Part1.　統計的手法による構文解析

Part2.　確率的文法モデリング

Part3.　確率的文法モデルの学習

Part4.　現在の到達点と今後の展開

Part1. 統計的手法による構文解析

自然言語処理における構文解析

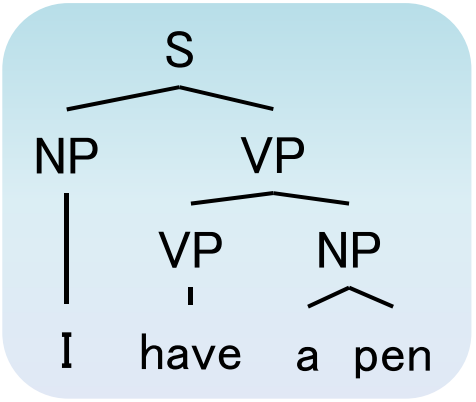
入力： 文

出力： 構文木

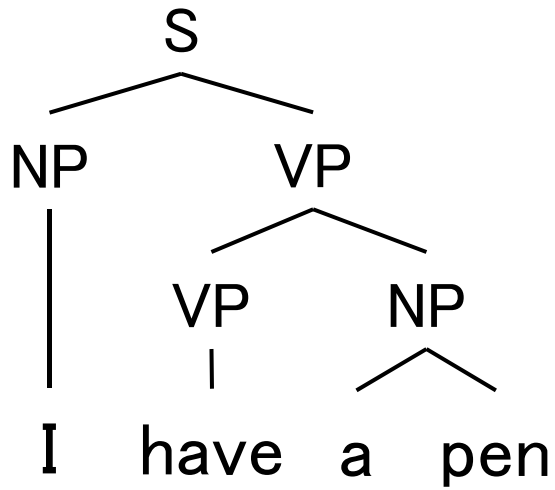
I have a pen

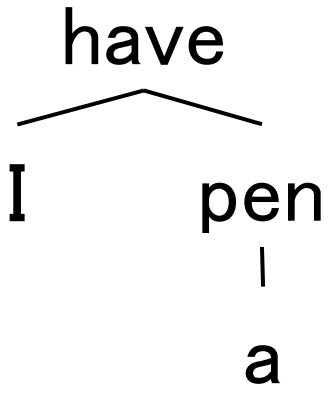
構文解析プログラム

統語構造



色々な種類の構文木がある





- 文脈自由文法
- 木置換文法
- 木接合文法
- 範疇文法
- 依存文法（係り受け）

文法の選択基準：

言語学的考慮　＋　計算機での扱いやすさ　＋　α

自然言語処理における構文解析

入力： 文

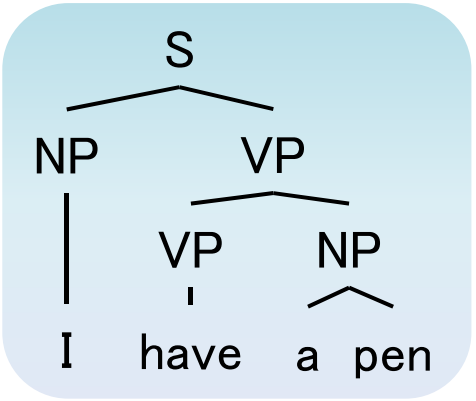
出力： 構文木

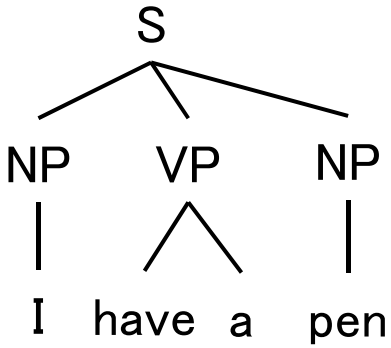
I have a pen

構文解析プログラム

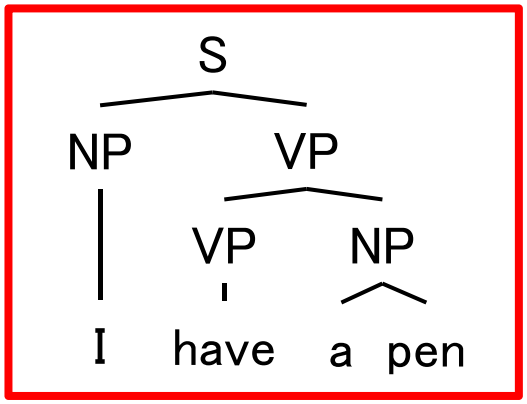
CYK法

統語構造

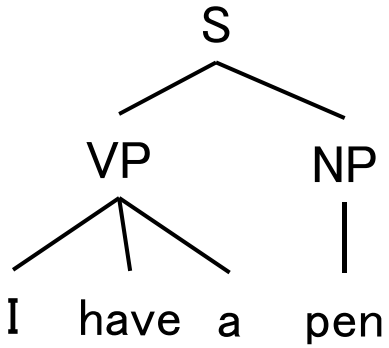




確率：0.001



確率：0.3



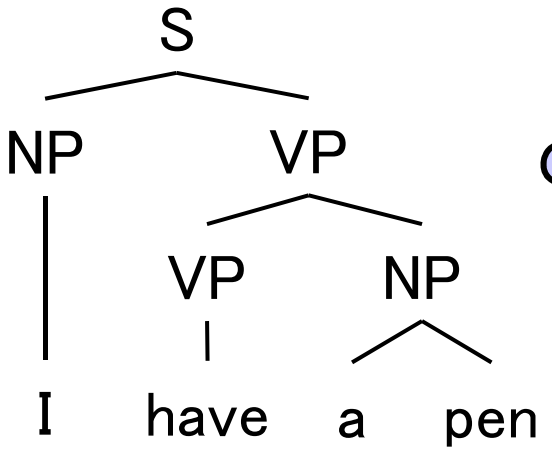
確率：0.02

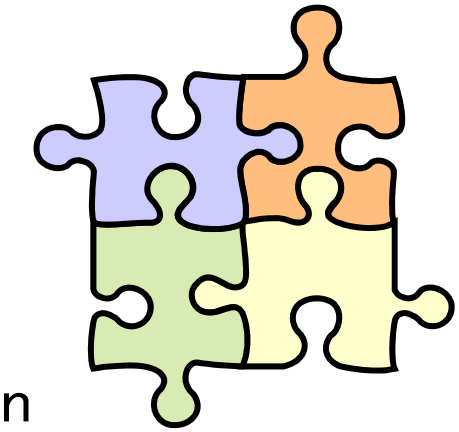
統計的手法による構文解析

P(構文木A) ＝ ？

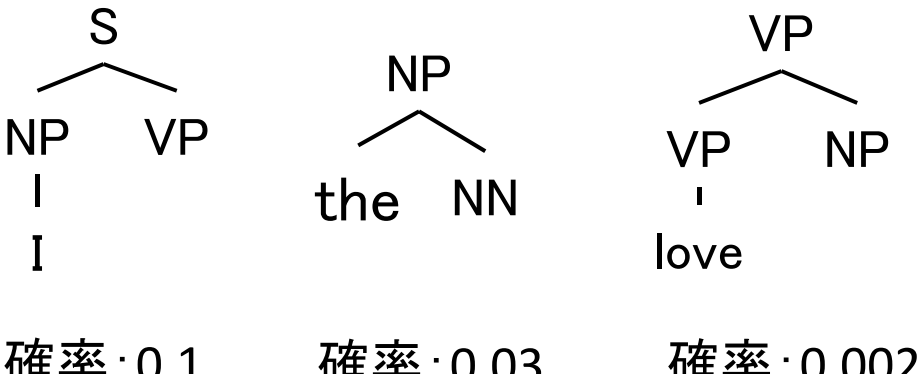
↓

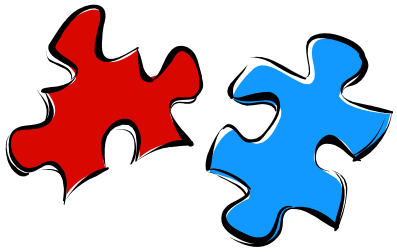
確率的文法モデル



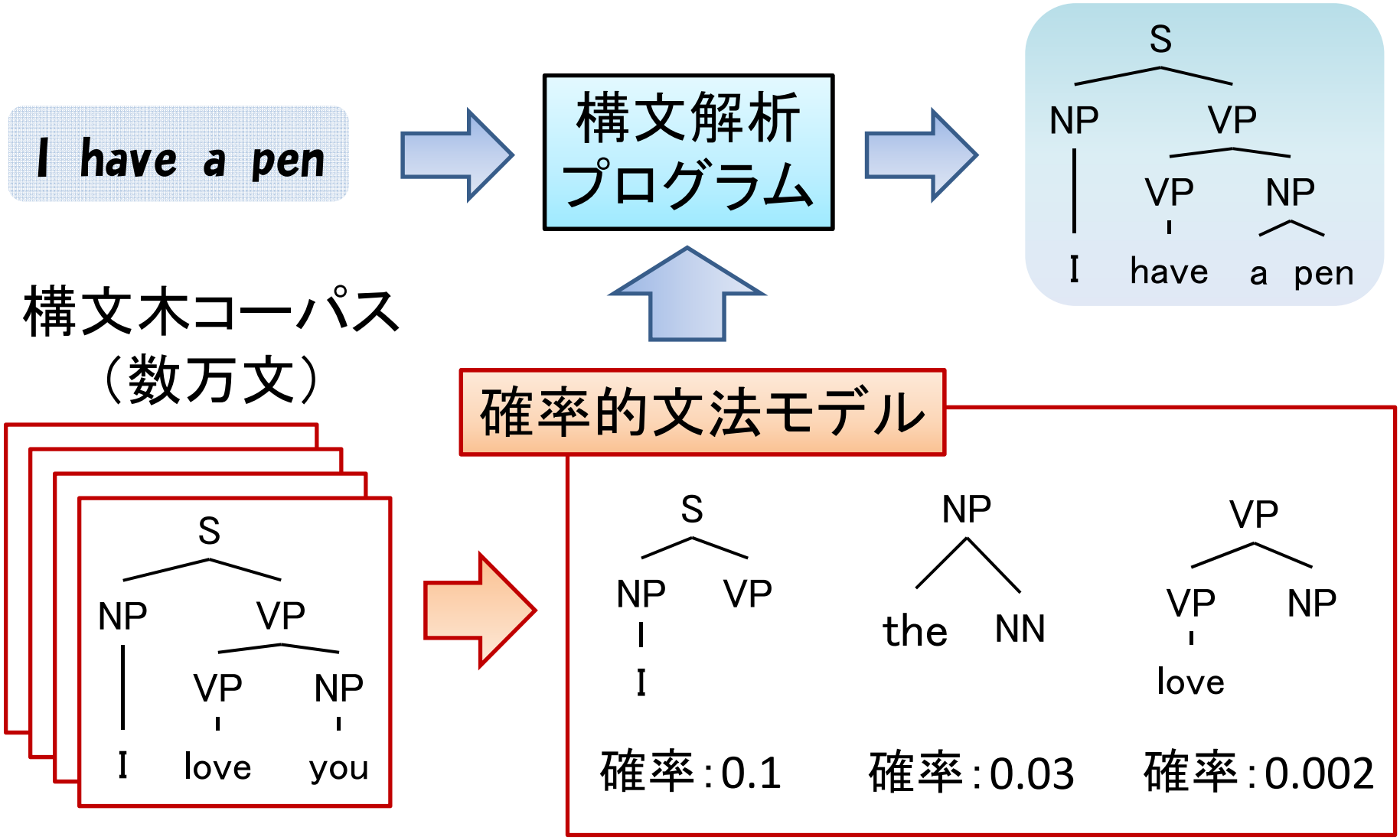


部分木を組み合わせて“P(構文木)”を計算する





統計的手法による構文解析



構文解析プログラムの作成へ向けて

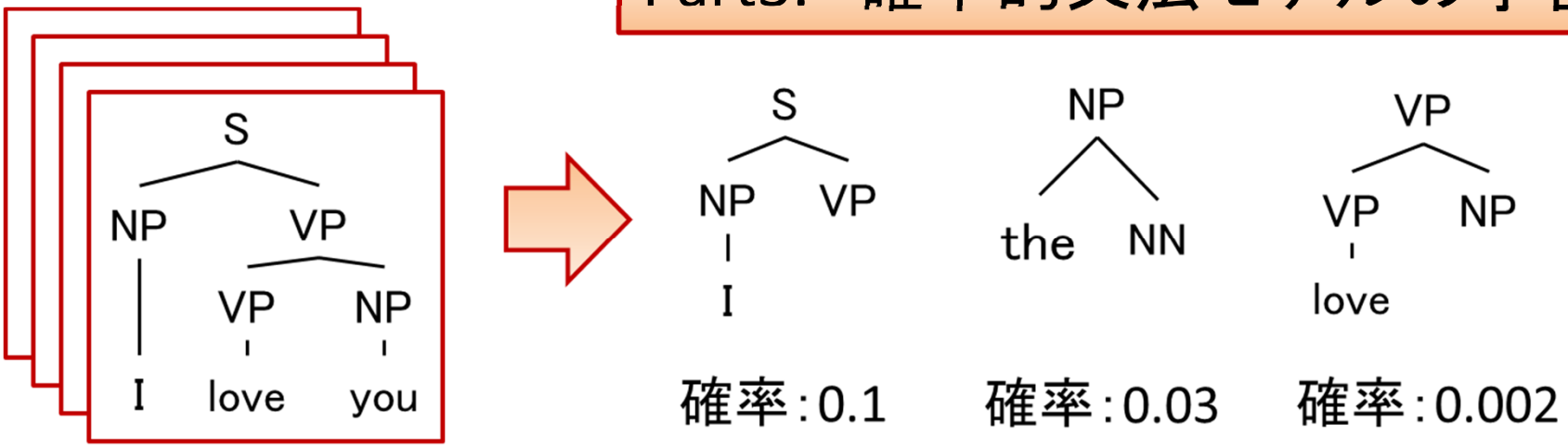
Q1.　構文木・部分木の確率を具体的に計算するには？

P（構文木）　＝　？

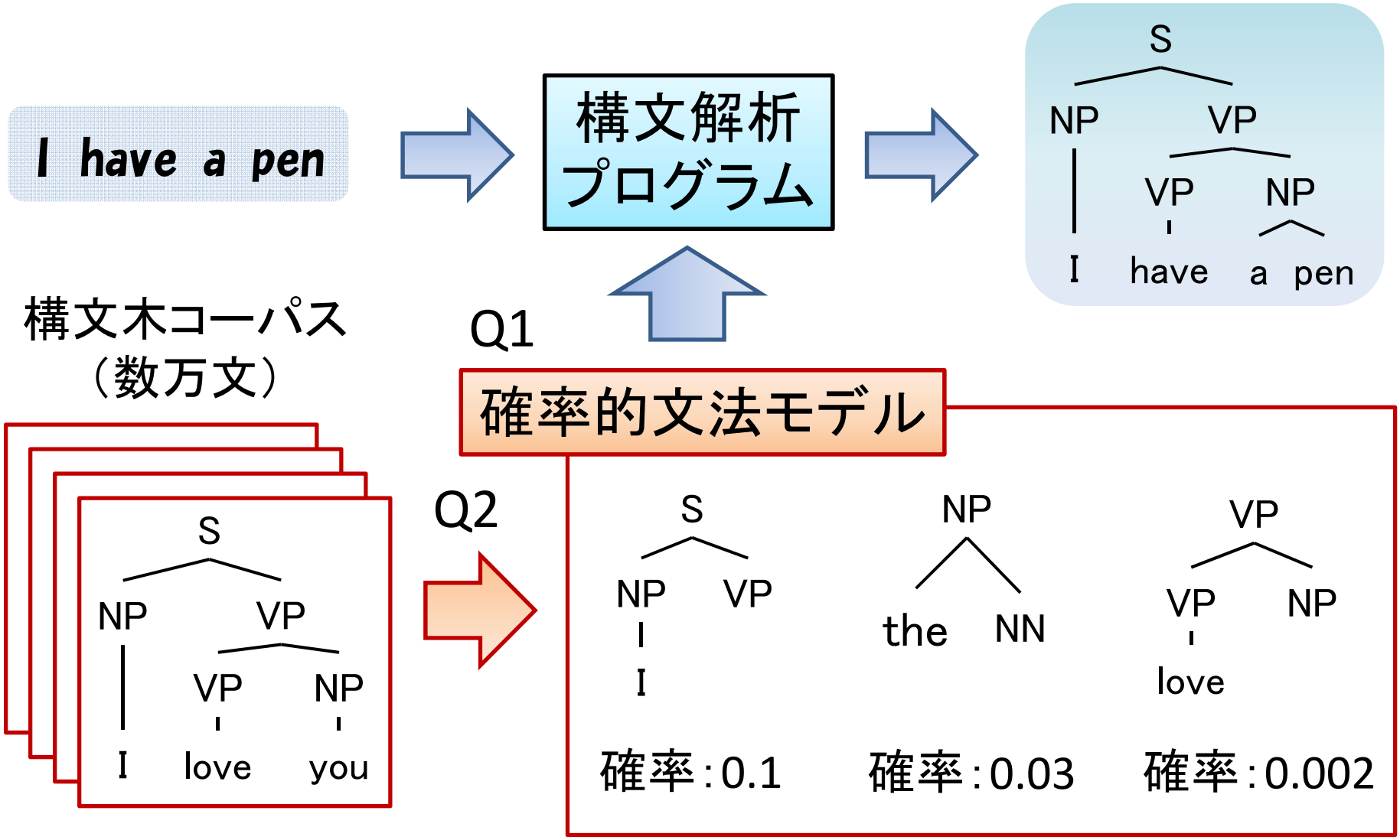
Part2.　確率的文法モデリング

Q2.　構文木コーパスから部分木を推定するには？

Part3.　確率的文法モデルの学習



統計的手法による構文解析



Part2.　確率的文法モデリング

確率的文法モデル

P(構文木) ＝ ？

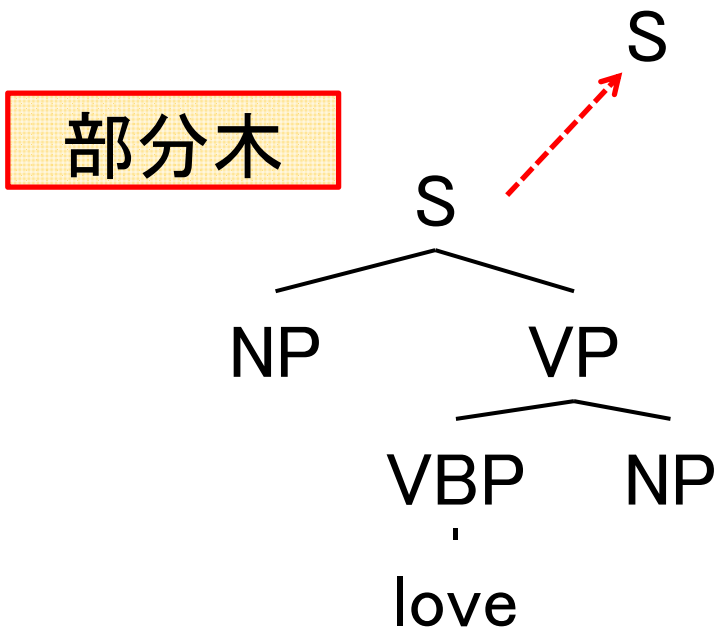
例1： 確率木置換文法

例2： 確率木接合文法

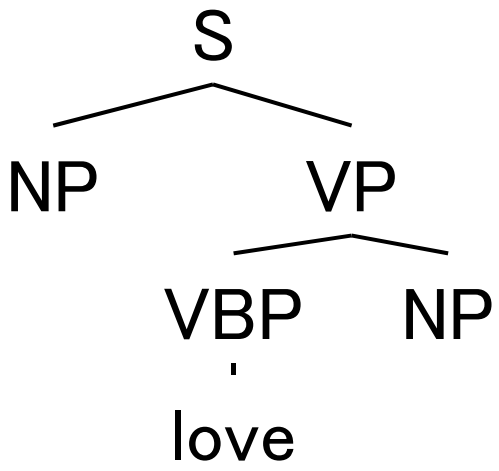
木置換文法

S

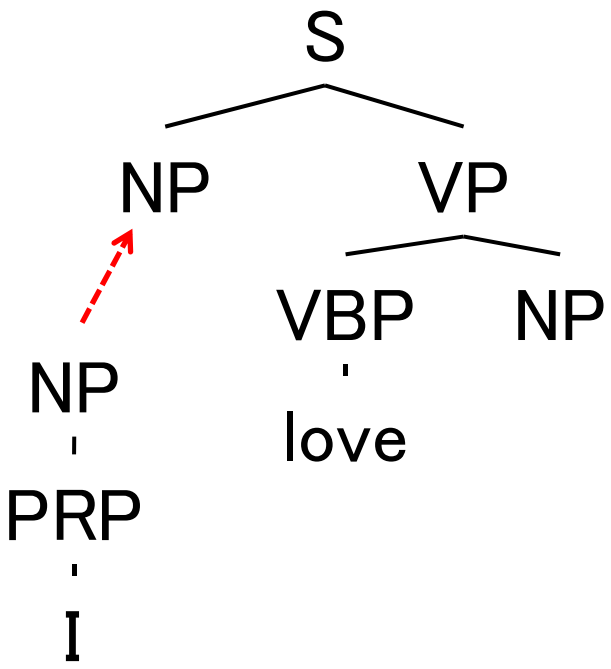
木置換文法



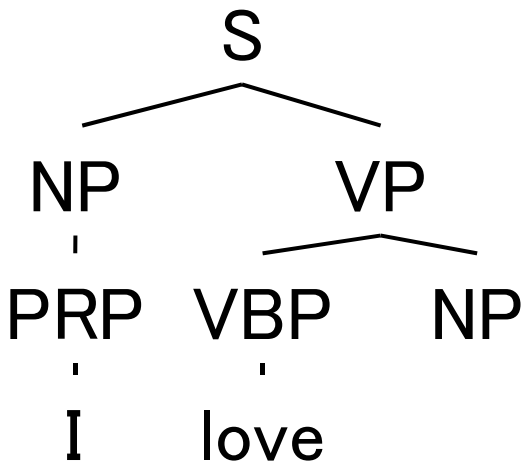
木置換文法



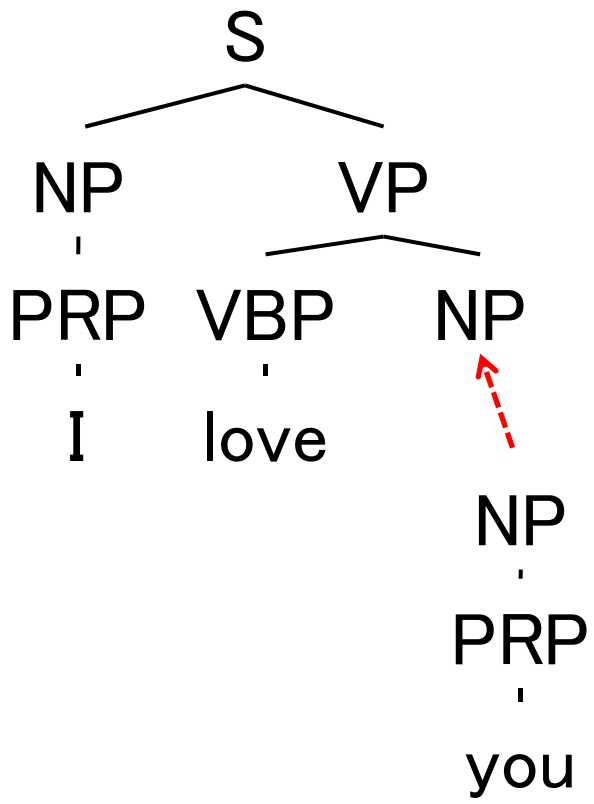
木置換文法



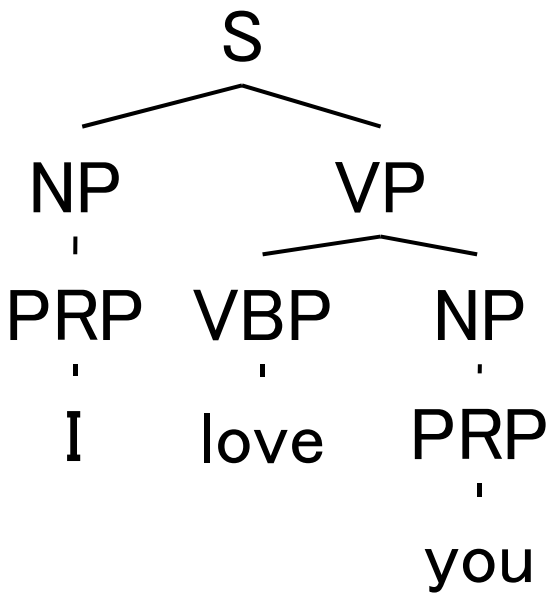
木置換文法



木置換文法

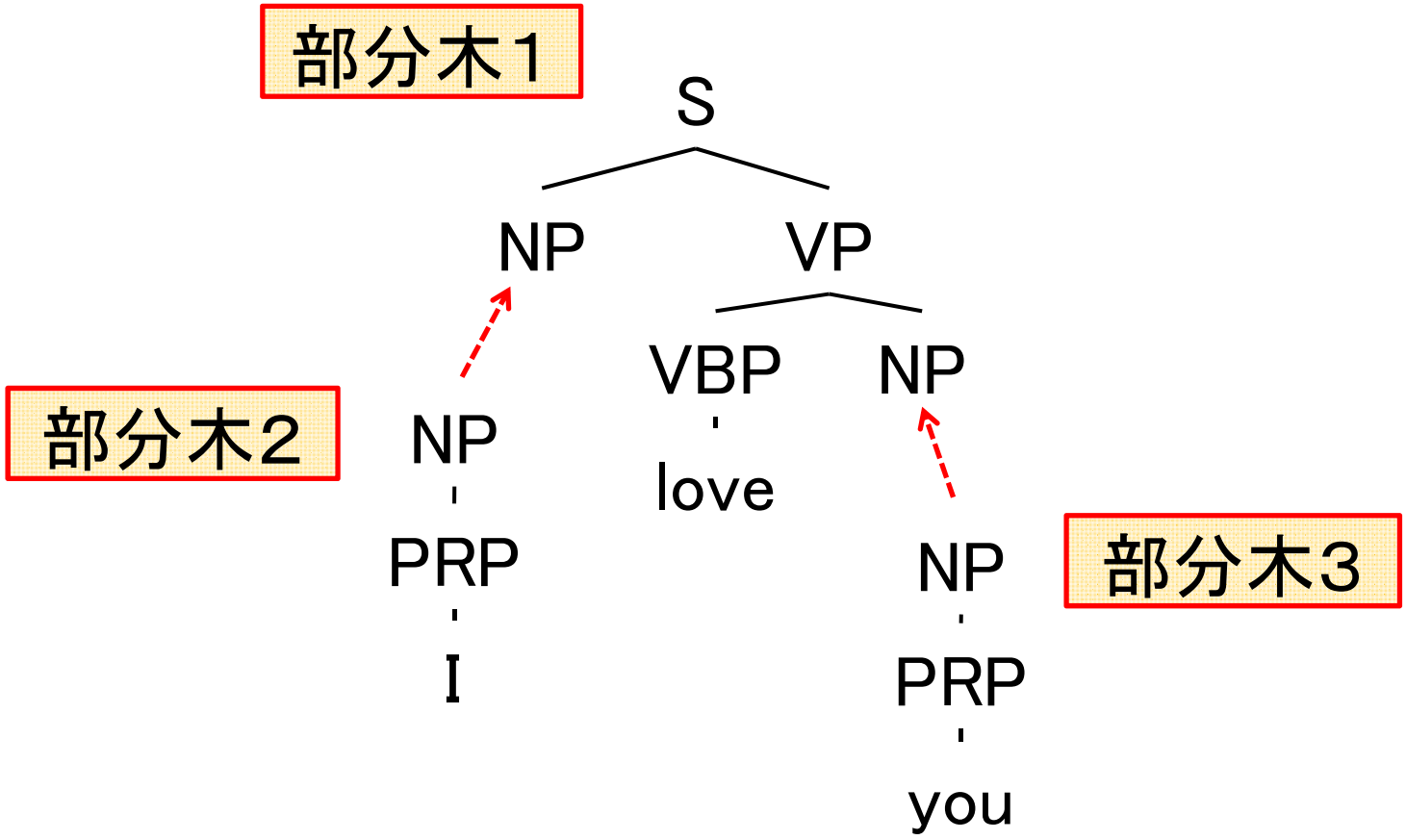


木置換文法



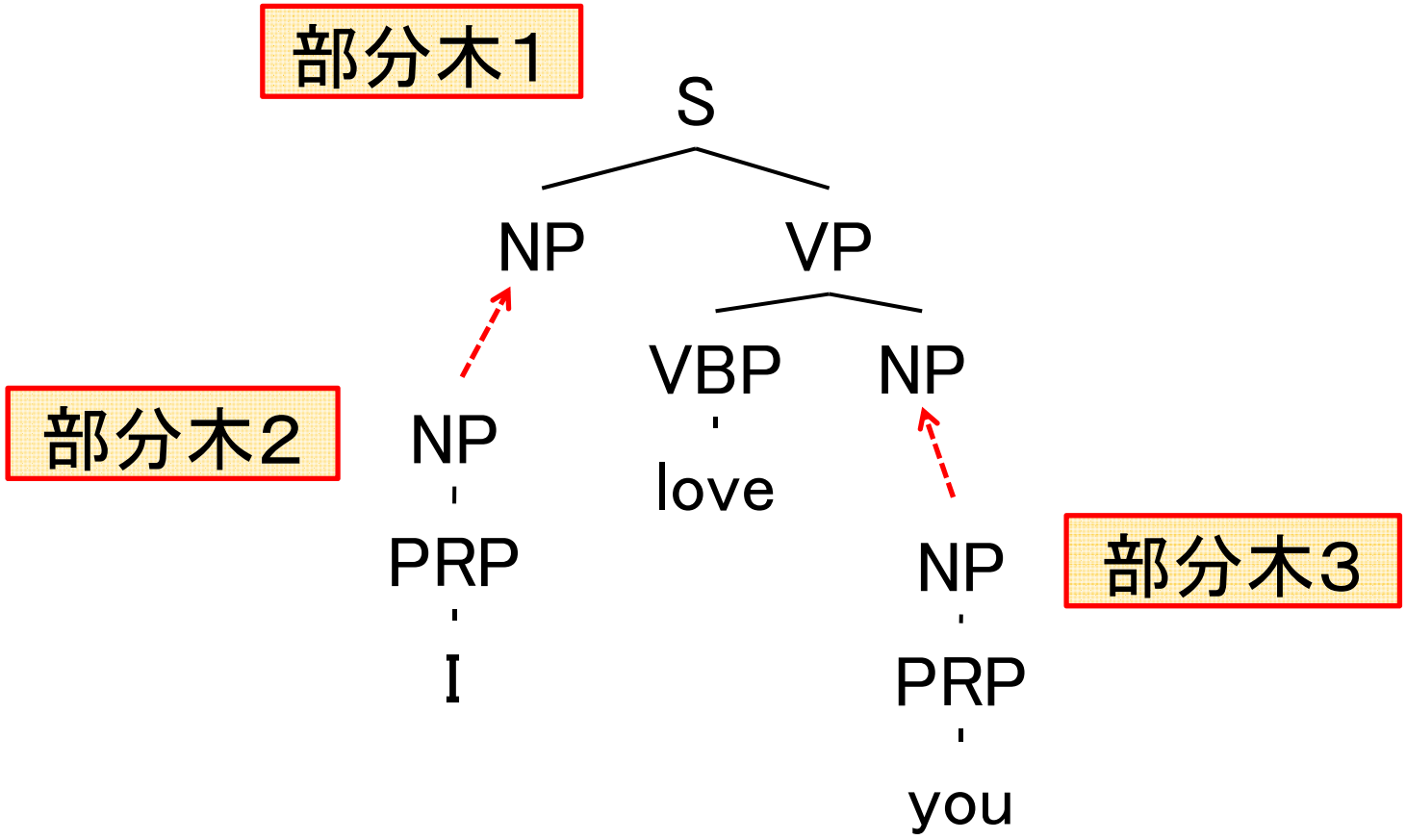
確率木置換文法

P(構文木)＝P(部分木1)×P(部分木2)×P(部分木3)



確率木置換文法

P(部分木)＝？

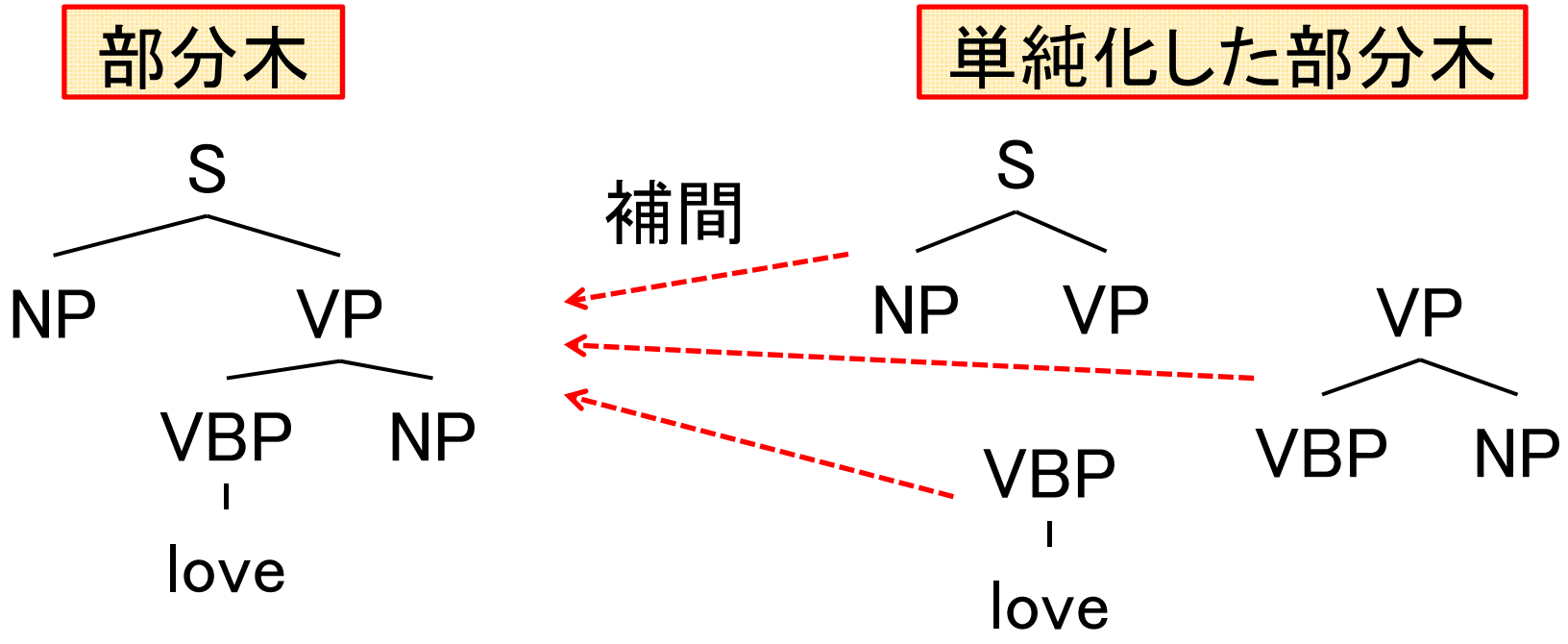


部分木の確率モデル

☹ $P_1(\text{部分木}) = \frac{1}{N} \times \text{count}(\text{部分木})$ ← 0になる可能性

☺ $P_2(\text{部分木}) = \alpha \times P_1(\text{部分木}) + \beta \times P_2(\text{単純化した部分木})$

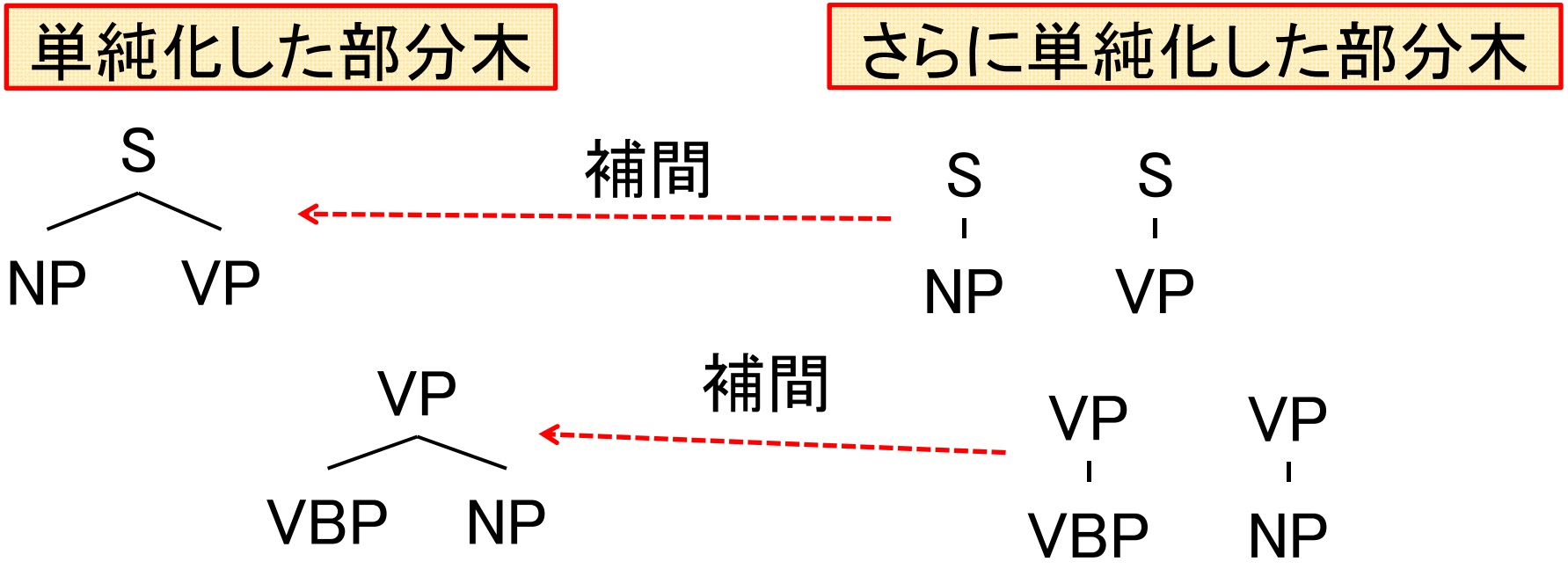
確率補間（スムージング）



部分木の確率モデル

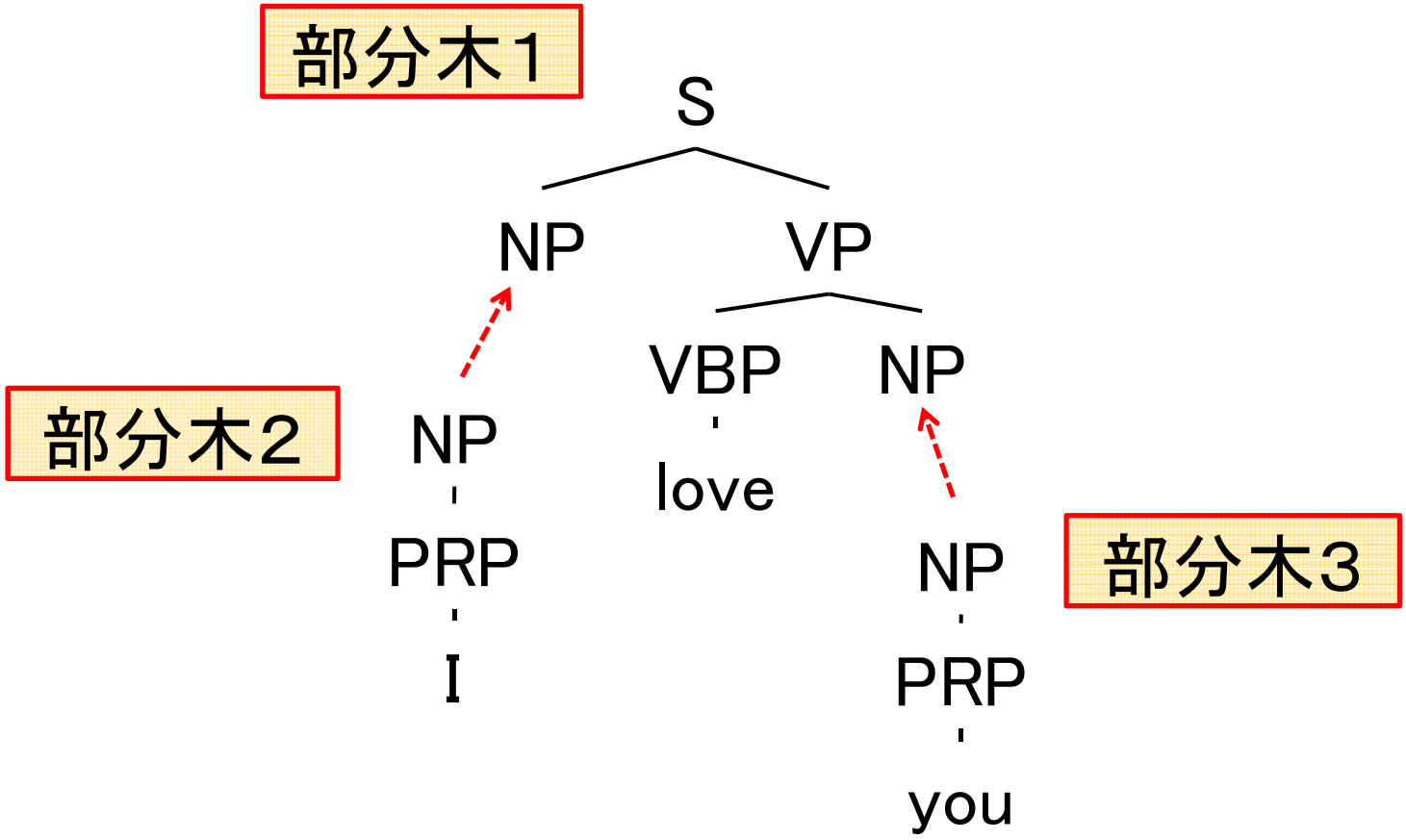
☹ $P_1(\text{部分木}) = \frac{1}{N} \times \text{count}(\text{部分木})$ ← 0になる可能性

☺ $P_2(\text{部分木}) = \alpha \times P_1(\text{部分木}) + \beta \times P_2(\text{単純化した部分木})$



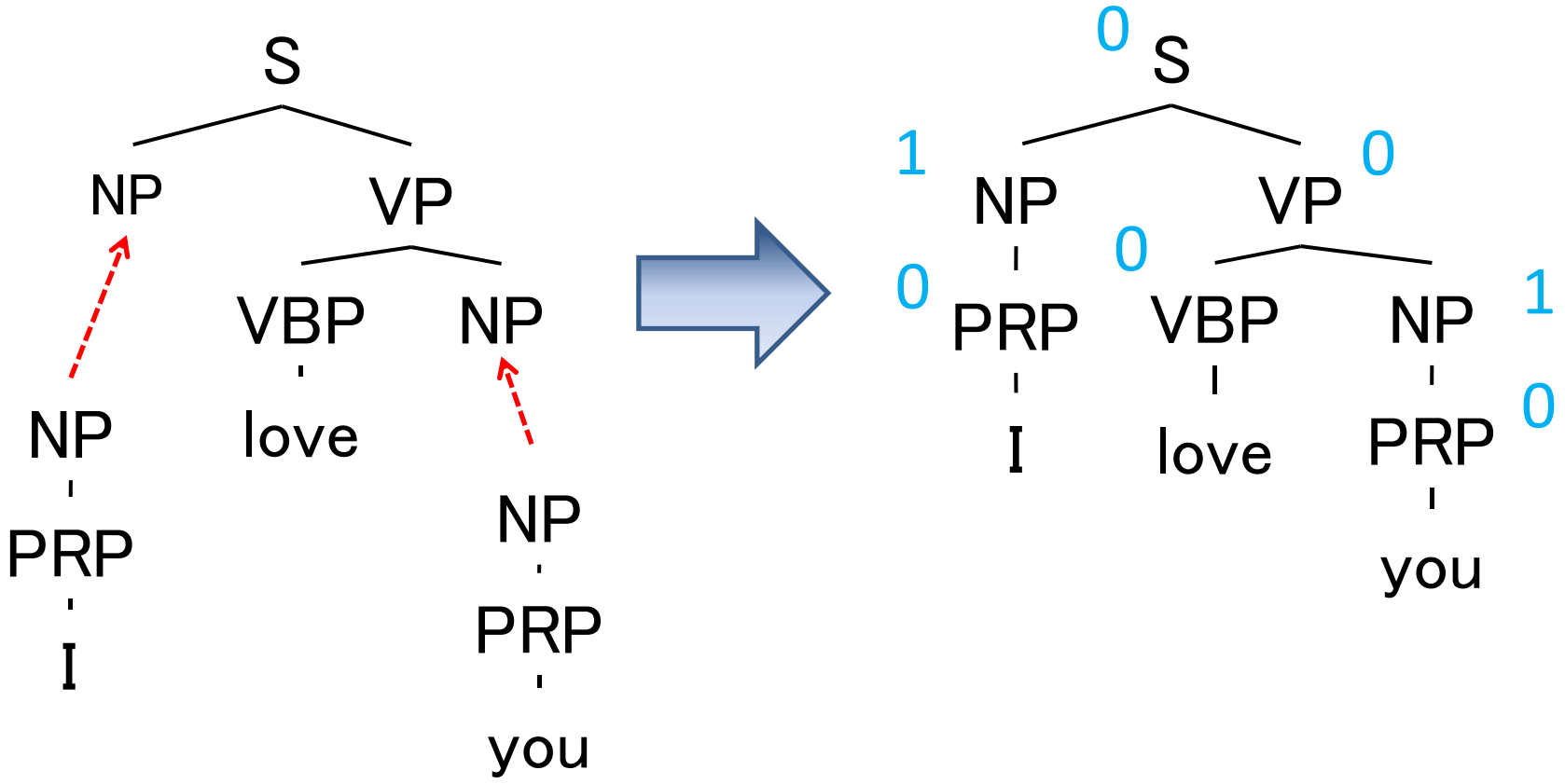
確率木置換文法

P(構文木)＝0.024



確率木置換文法

木置換文法の情報をノードに埋め込む



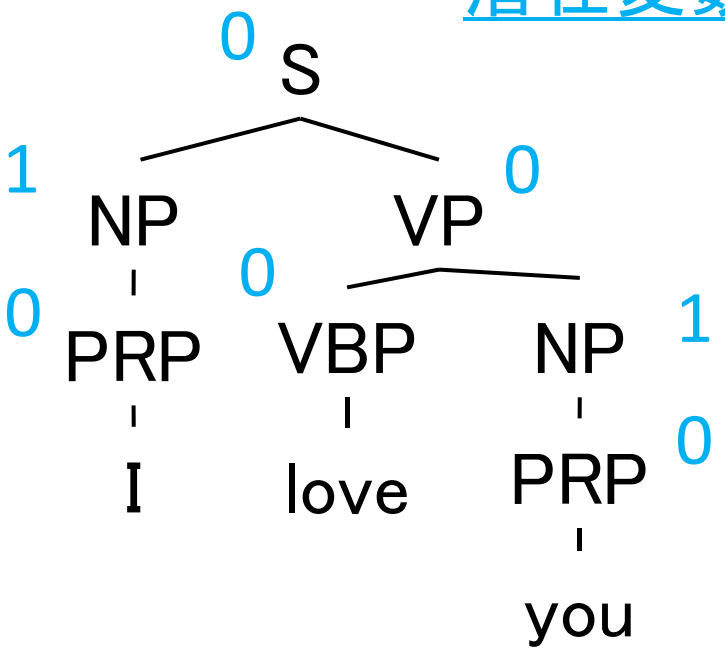
潜在変数を含む確率モデルになる

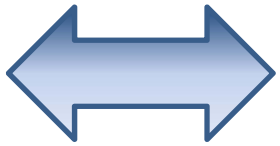
確率的木置換文法

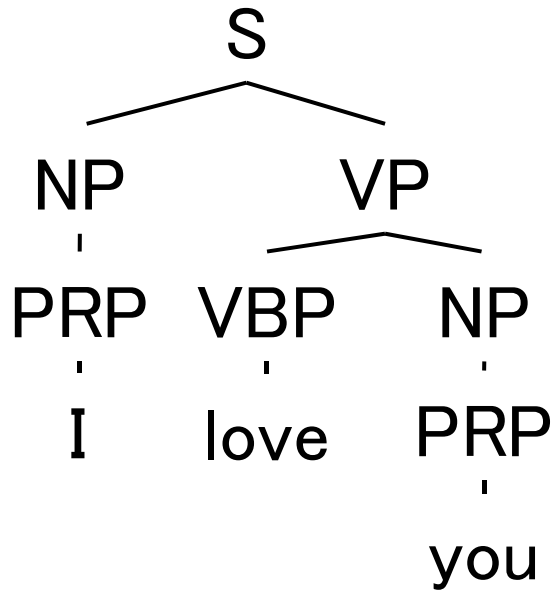
構文木

＝　観測データ

潜在変数







例１：　確率木置換文法

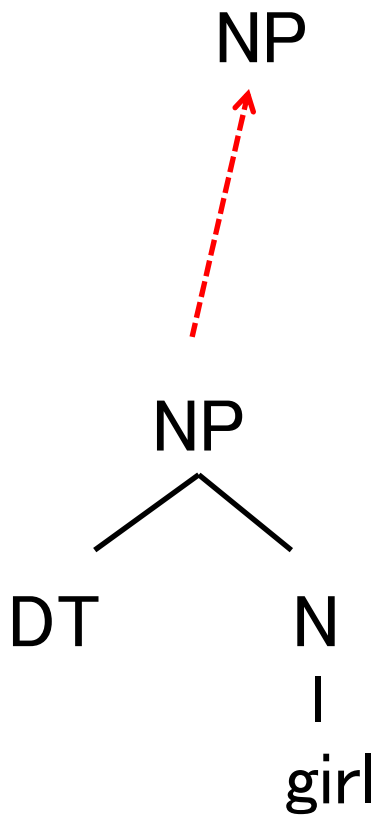
例２：　確率木接合文法

＝　確率木置換文法　＋　部分木の挿入操作

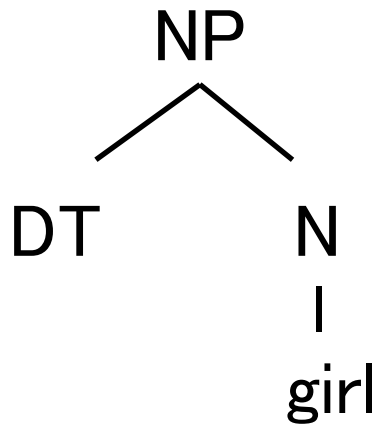
確率木接合文法

NP

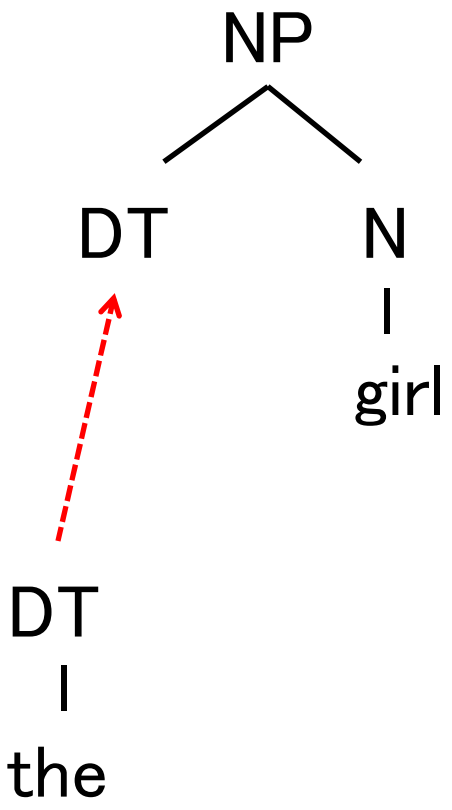
確率木接合文法



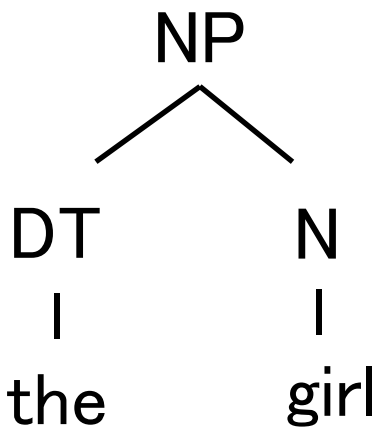
確率木接合文法



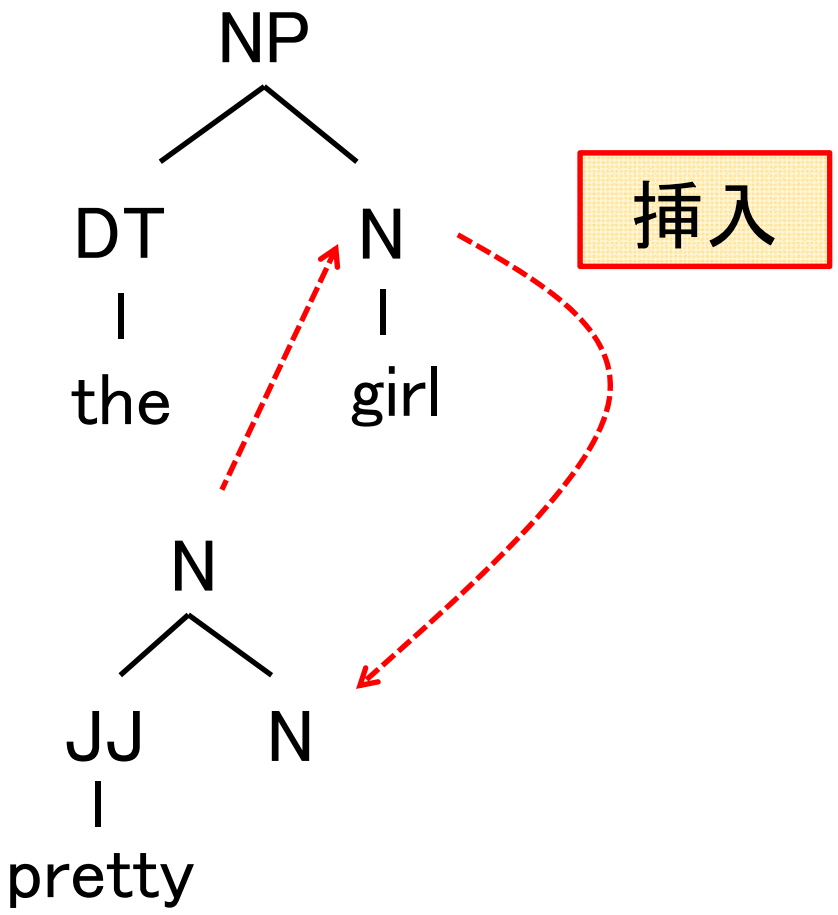
確率木接合文法



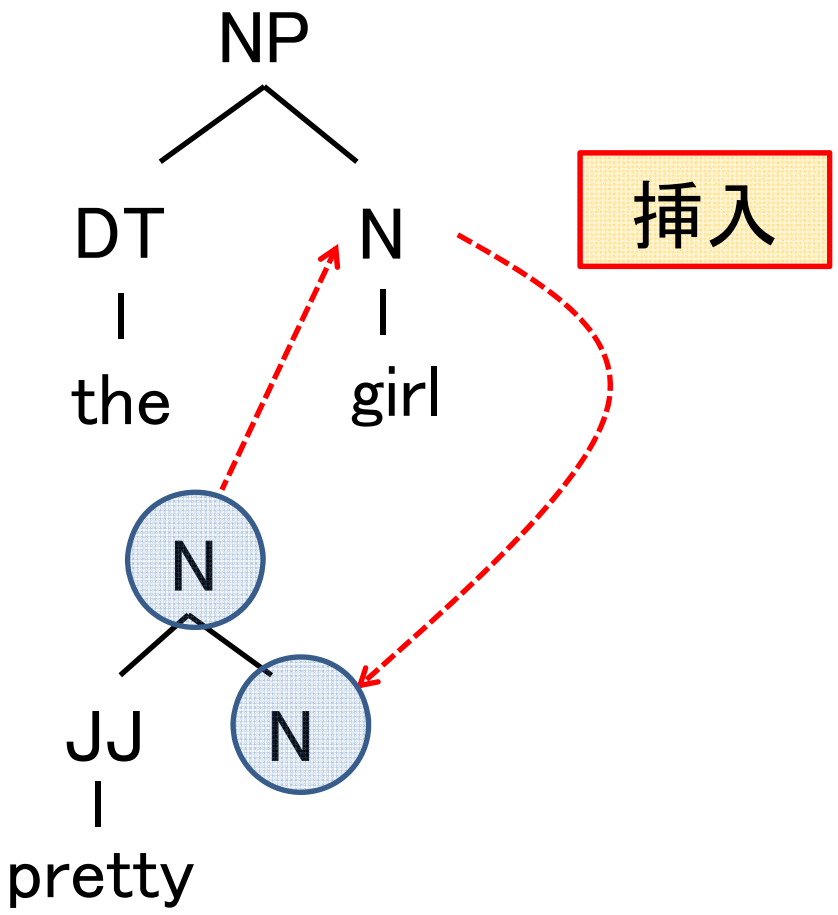
確率木接合文法



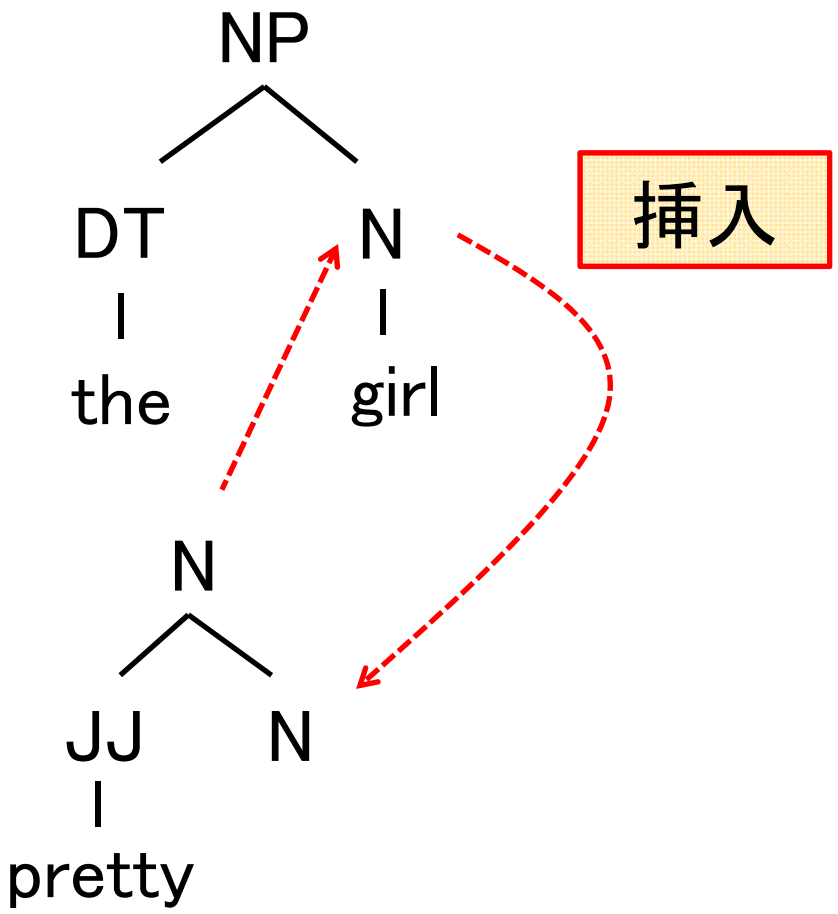
確率木接合文法



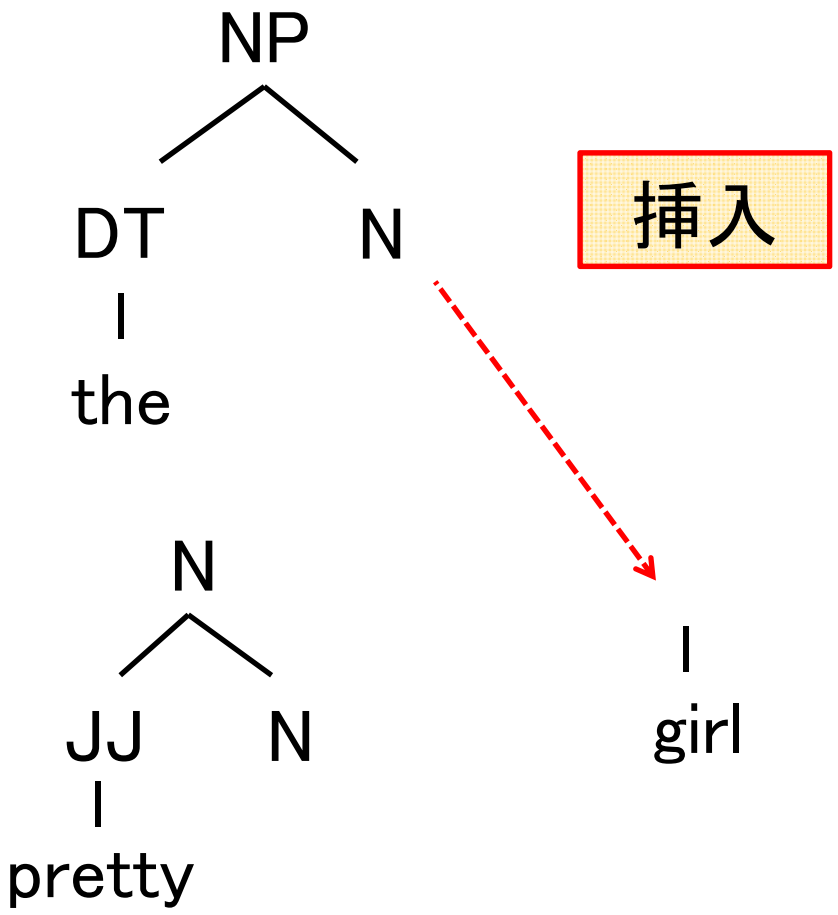
確率木接合文法



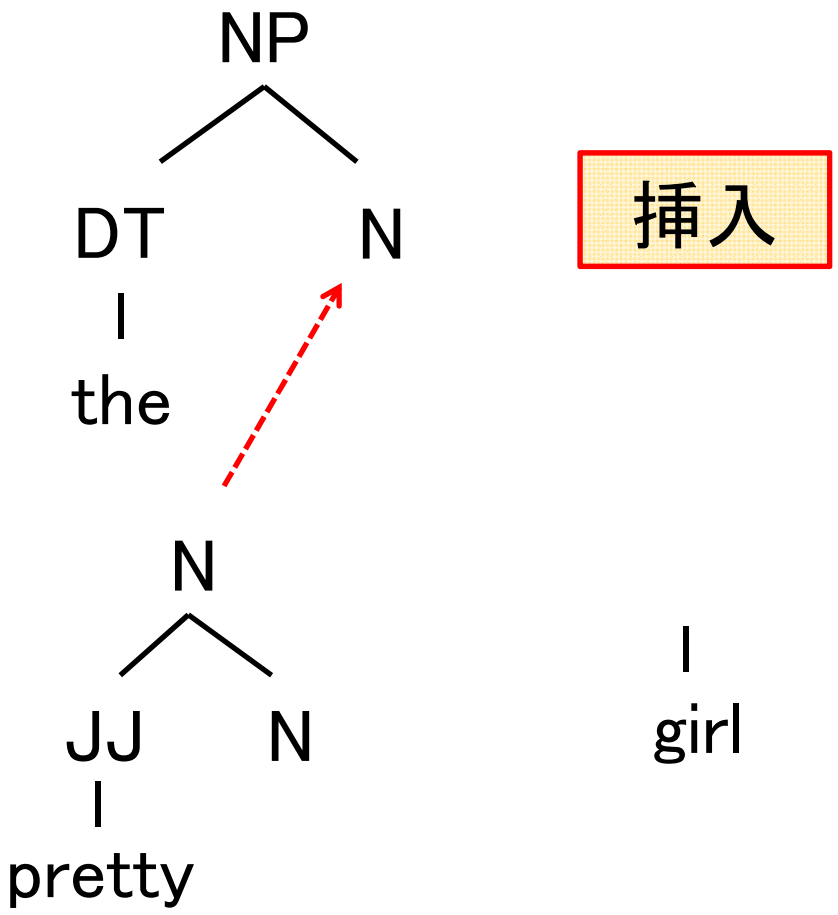
確率木接合文法



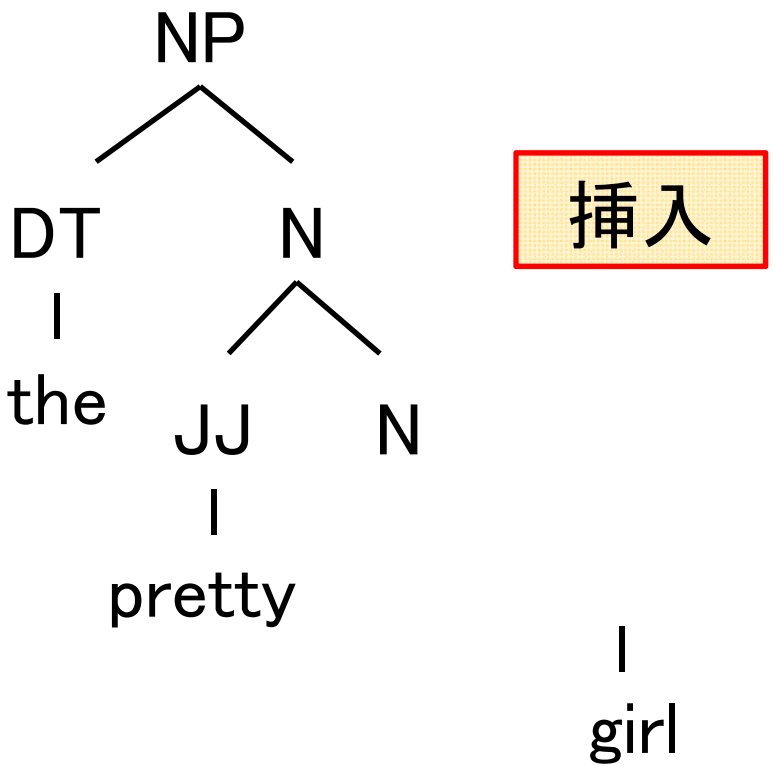
確率木接合文法



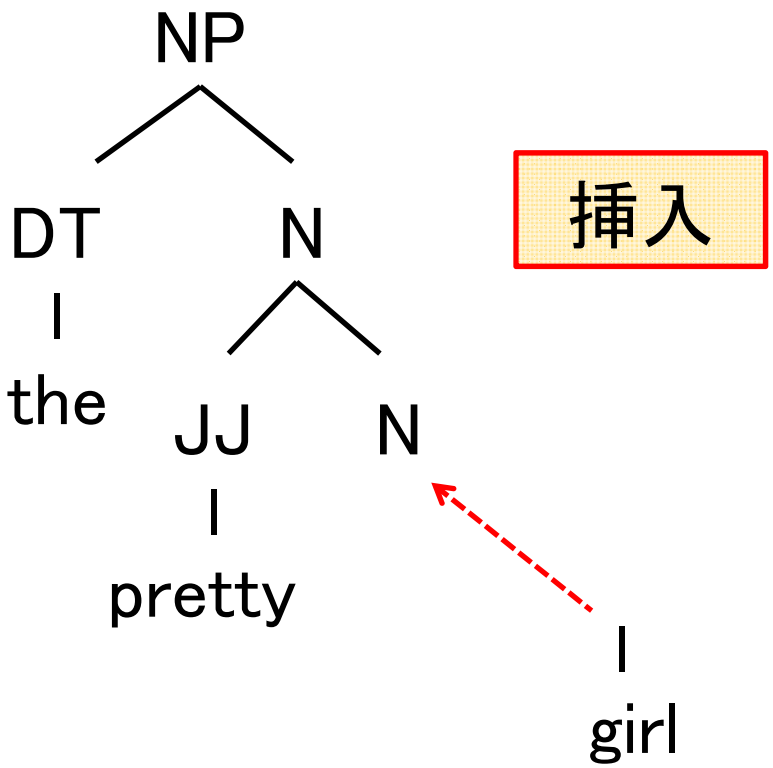
確率木接合文法



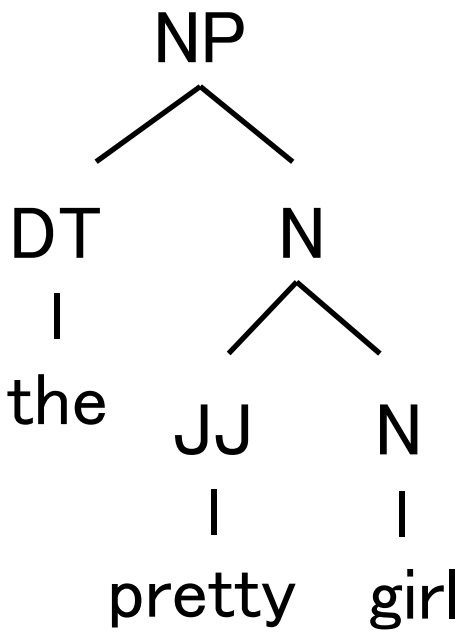
確率木接合文法



確率木接合文法



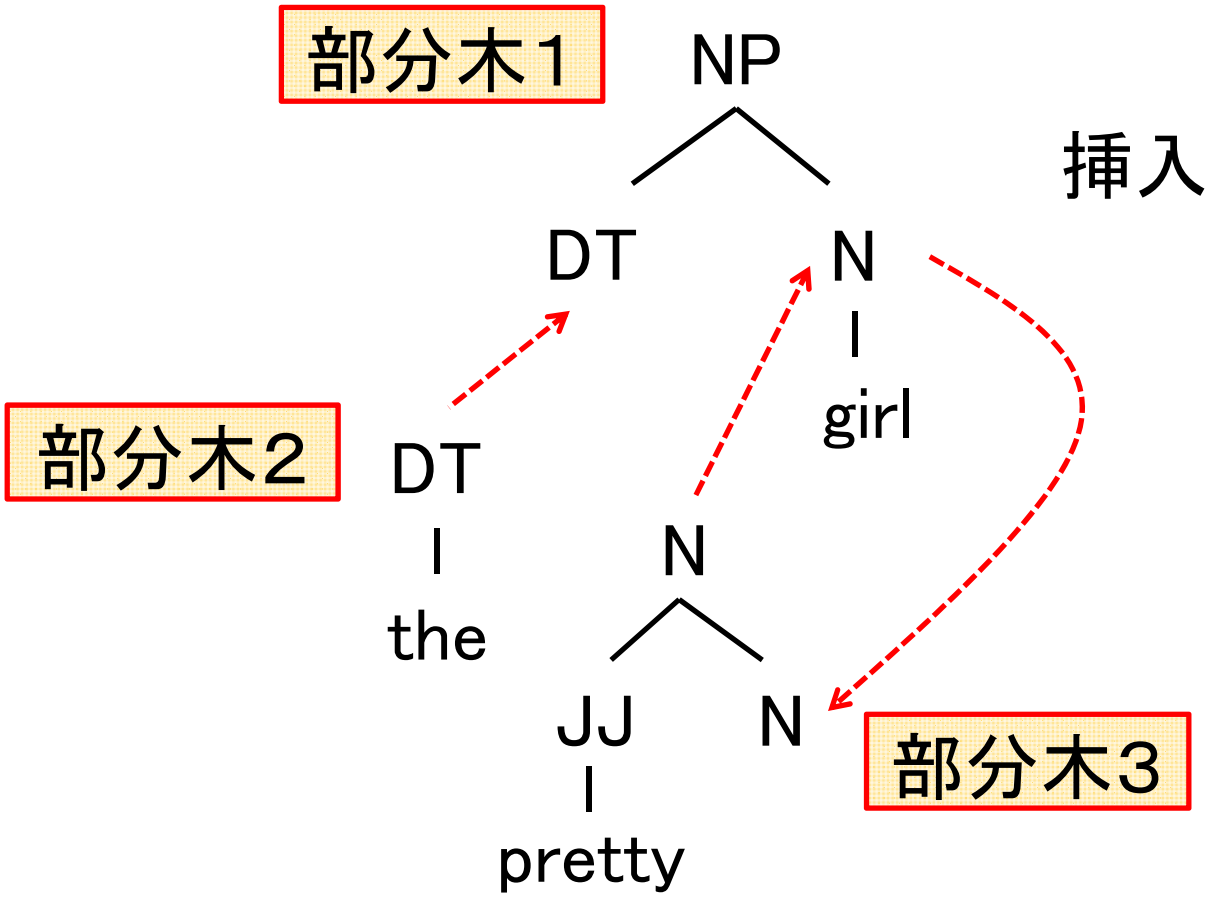
確率木接合文法



確率木接合文法

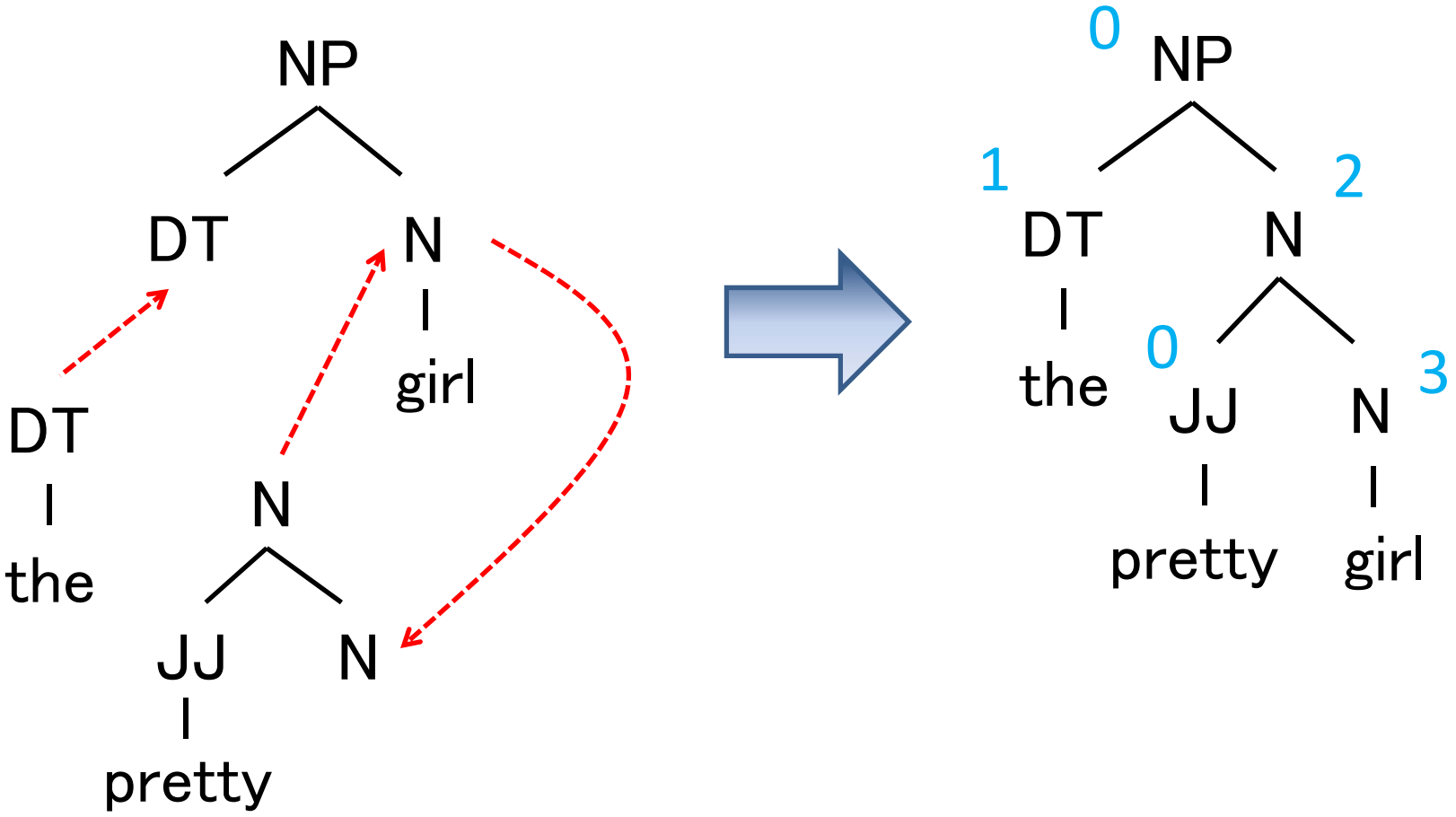
P(構文木)＝P(部分木1)×P(部分木2)×P'(部分木3)

挿入用の確率分布



確率木接合文法

木接合文法の情報をノードに埋め込む



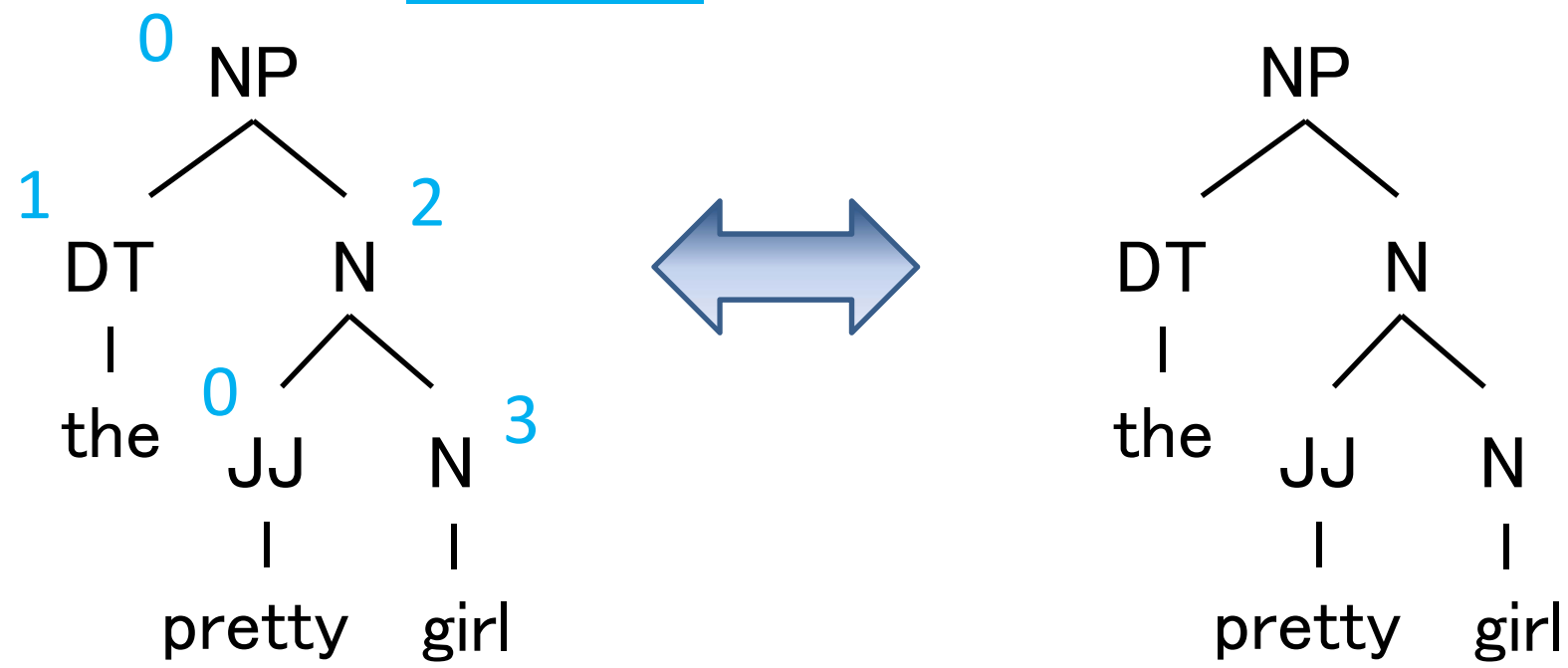
潜在変数を含む確率モデルになる

確率木接合文法

構文木

＝　観測データ

潜在変数



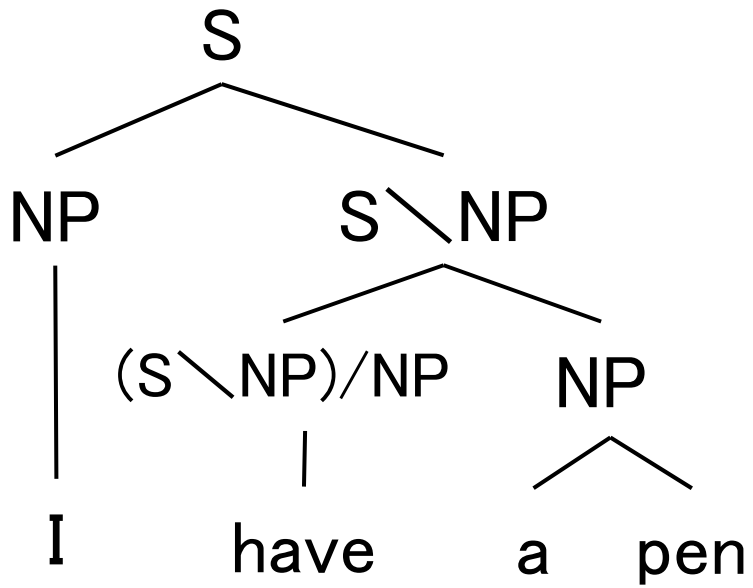
例1： 確率木置換文法

例2： 確率木接合文法

例3： 確率範疇文法

確率範疇文法

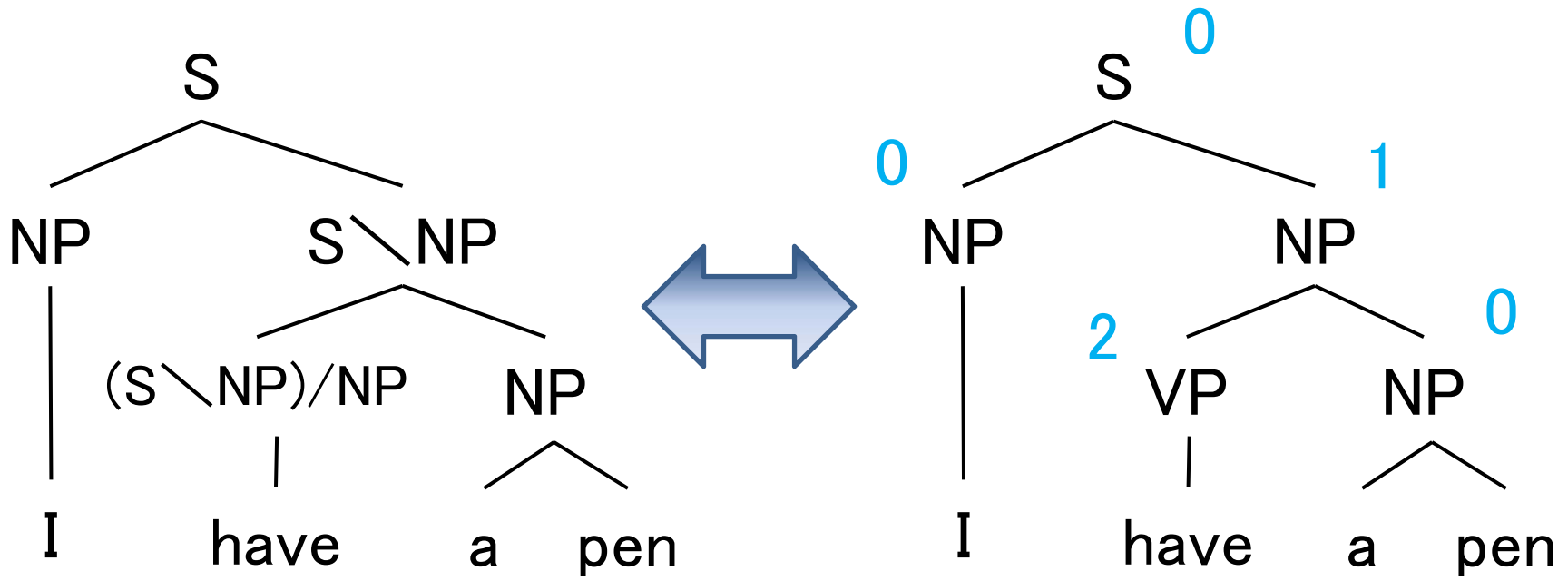
P(構文木)＝P(部分木1)×P(部分木2)×P(部分木3)



P(部分木)：　対数線形モデルなど

確率範疇文法

範疇文法の情報をノードに埋め込む



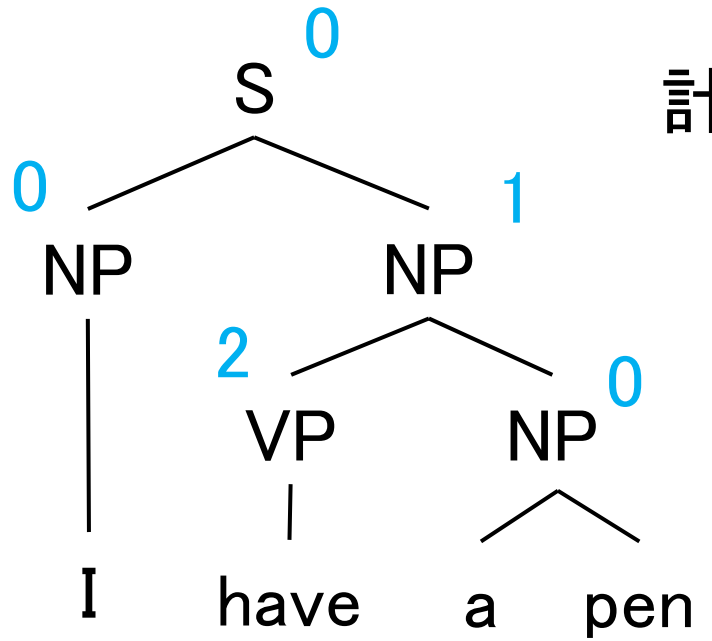
潜在変数を含む確率モデルになる

確率範疇文法

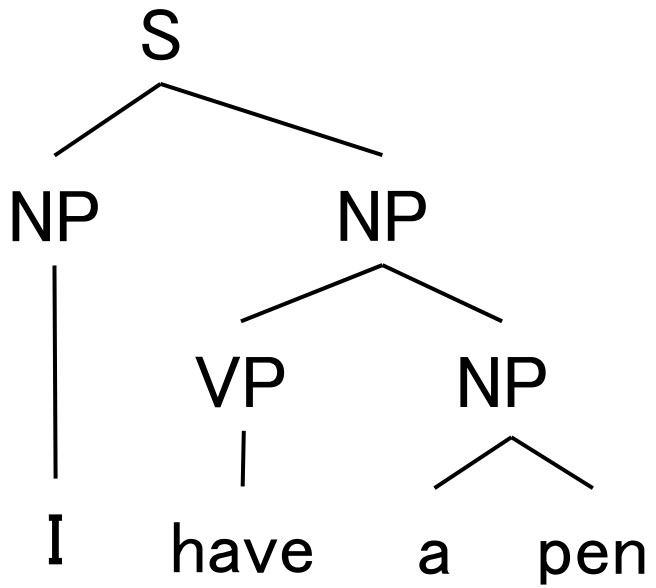
構文木

＝　観測データ

潜在変数



計算機が推定



例１：　確率木置換文法

例２：　確率木接合文法

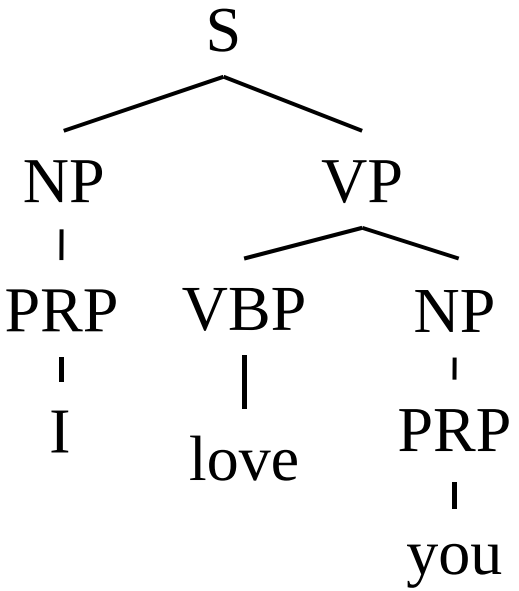
例３：　確率範疇文法

例４：　確率Ｘ文法　→　同様に潜在変数モデル化可能

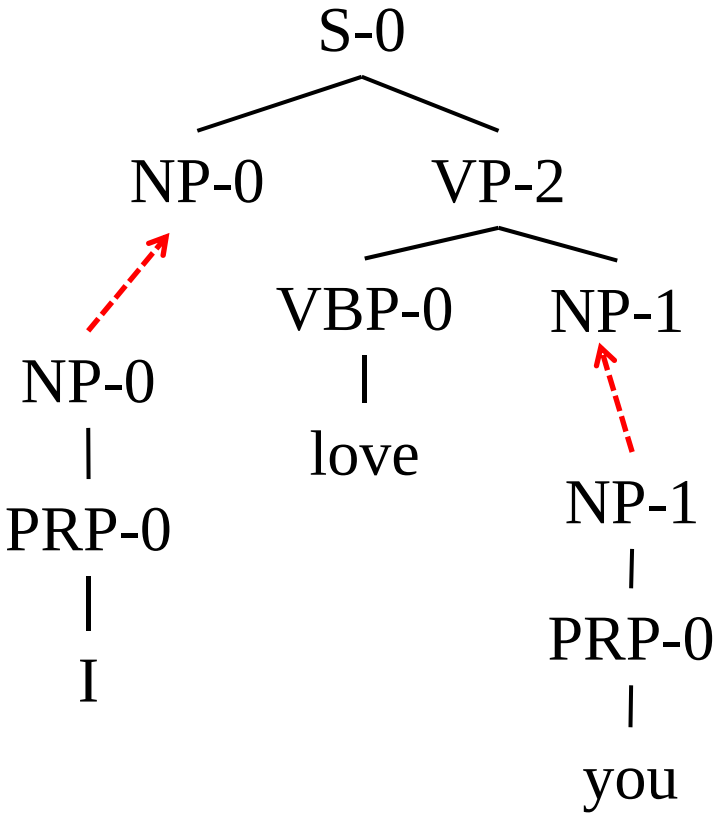
(参考)シンボル細分化木置換文法

木置換文法　＋　シンボル細分化

構文木



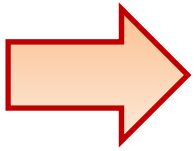
SR-TSG



[shindo et al. ACL 2012]

Part2. まとめ

1. P(構文木)＝P(部分木1)×P(部分木2)×……
2. 文法モデルを選ぶ／自分で作る
3. 構文木に含まれていない情報　→　潜在変数
4. 潜在変数の推定は計算機に任せる



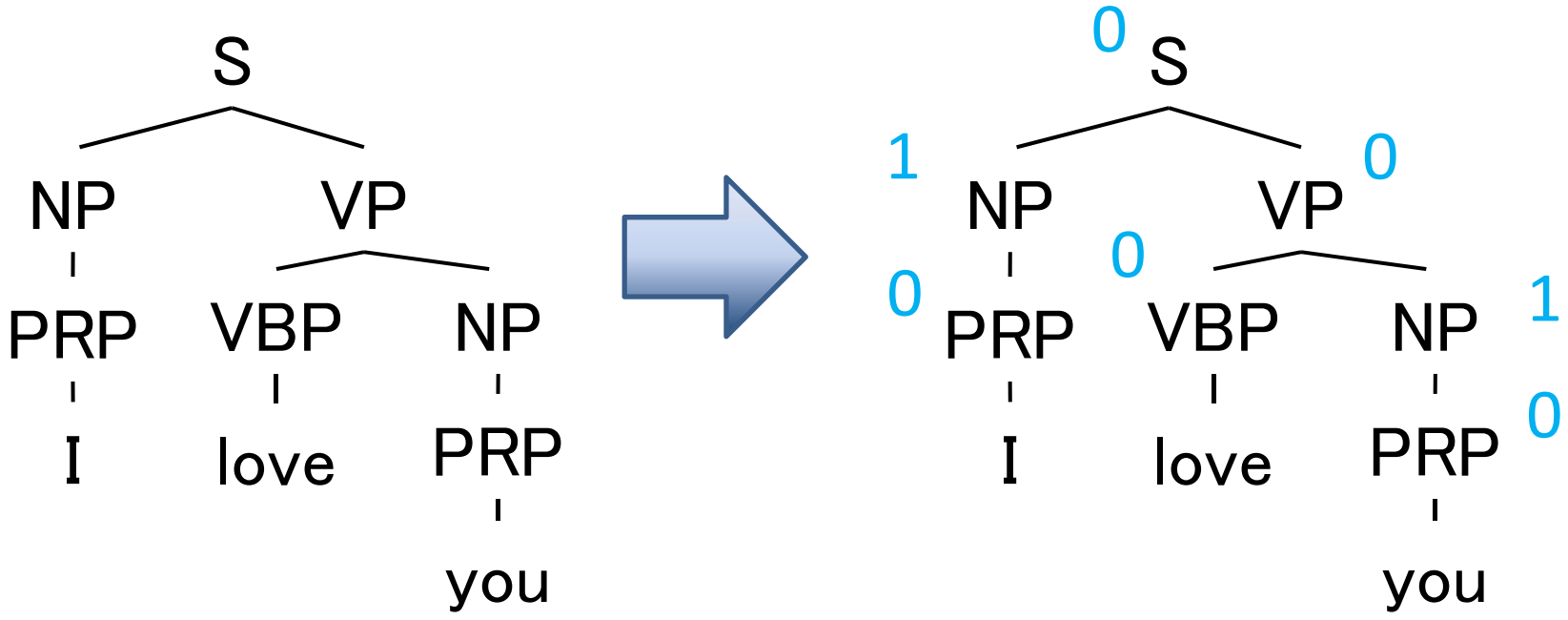
Part3. 確率的文法モデルの学習

Part3. 確率的文法モデルの学習

確率的文法モデルの学習

P(構文木)を最大にする潜在変数を求める

P(構文木)＝P(部分木1)×P(部分木2)×P(部分木3)

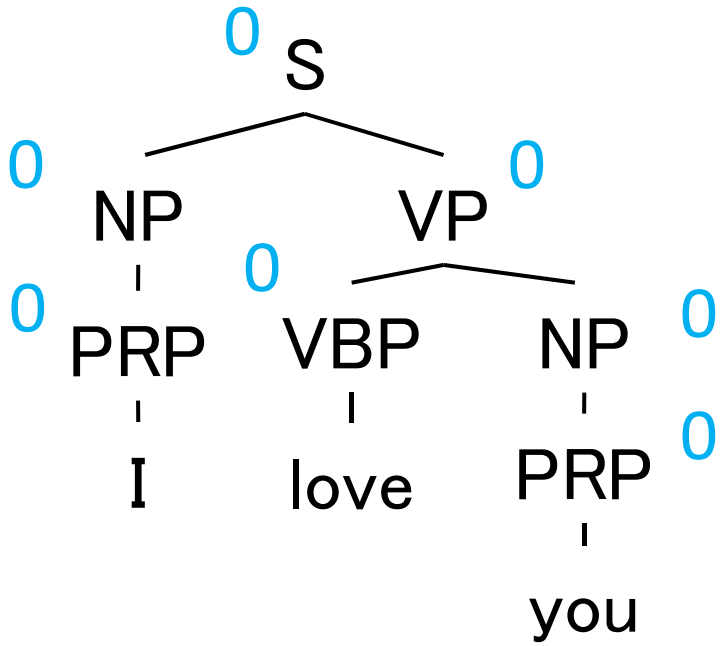


確率的文法モデルの学習

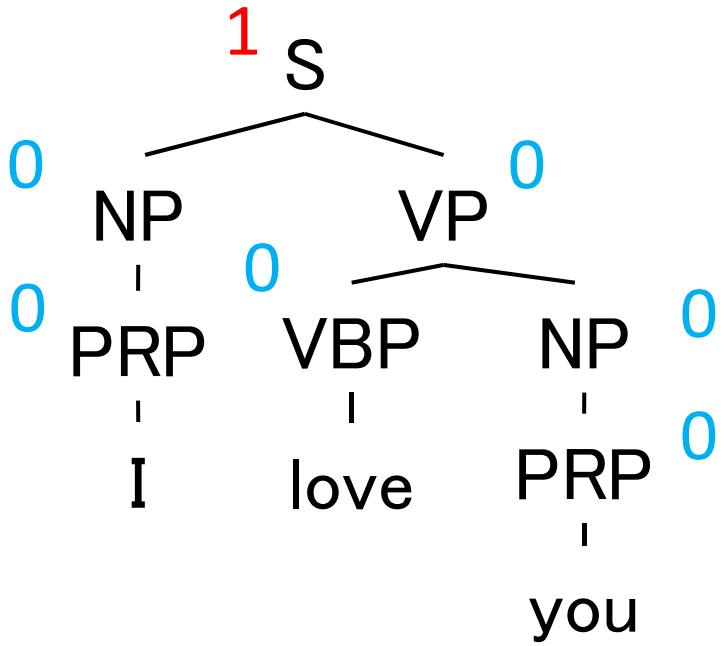
木構造データの潜在変数を推定する方法

- マルコフ連鎖モンテカルロ法 [Johnson 07]
- 期待値最大化（EM）法 [Matsuzaki 05] [Petrov 06]
- 変分ベイズ法 [Liang 07] [Coehn 10]

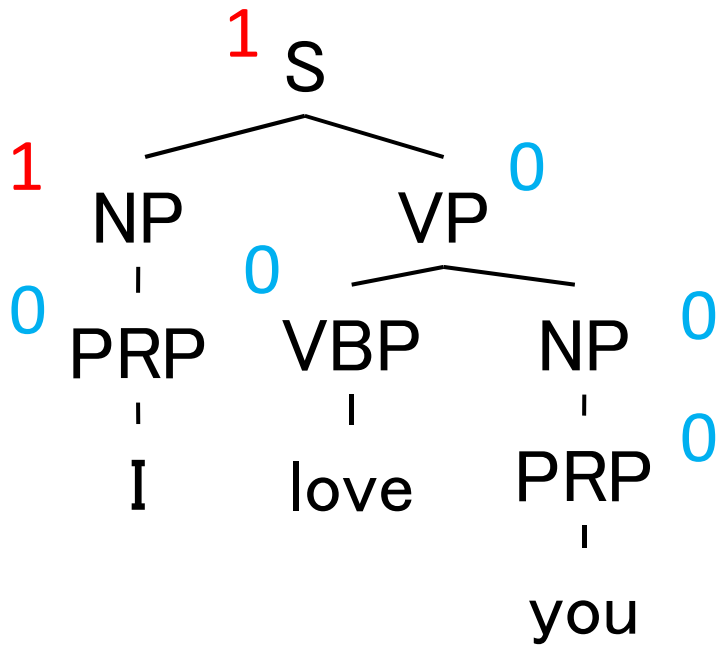
ギブスサンプリング



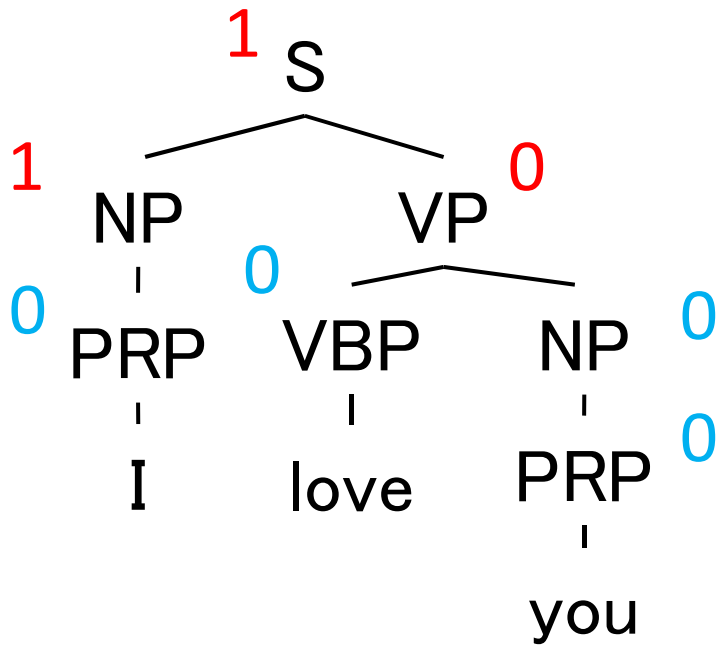
ギブスサンプリング



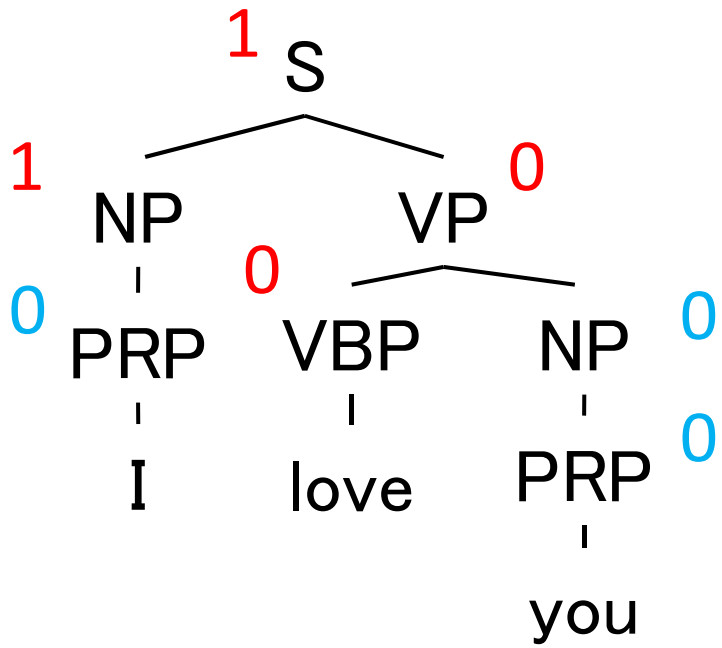
ギブスサンプリング



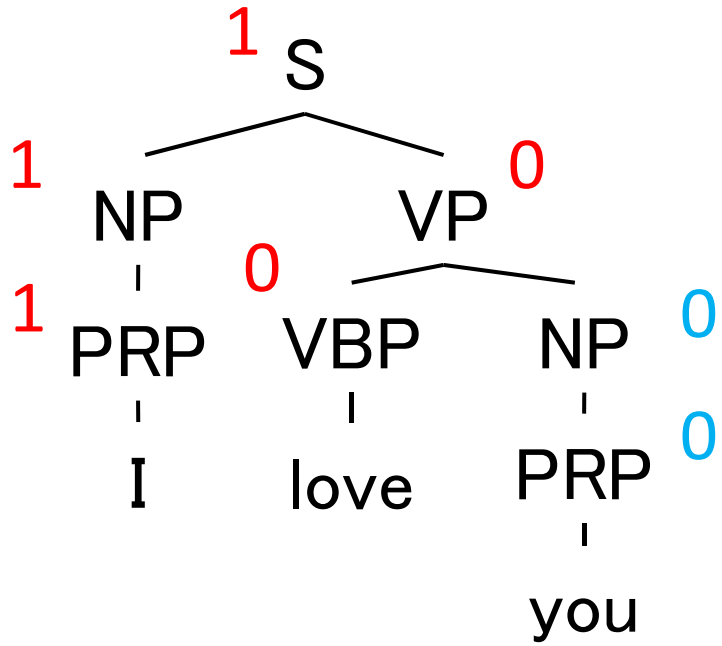
ギブスサンプリング



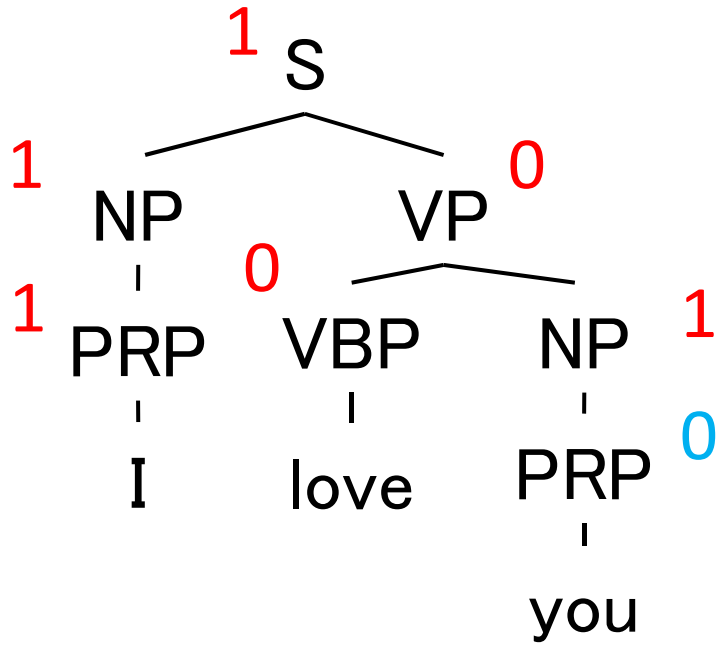
ギブスサンプリング



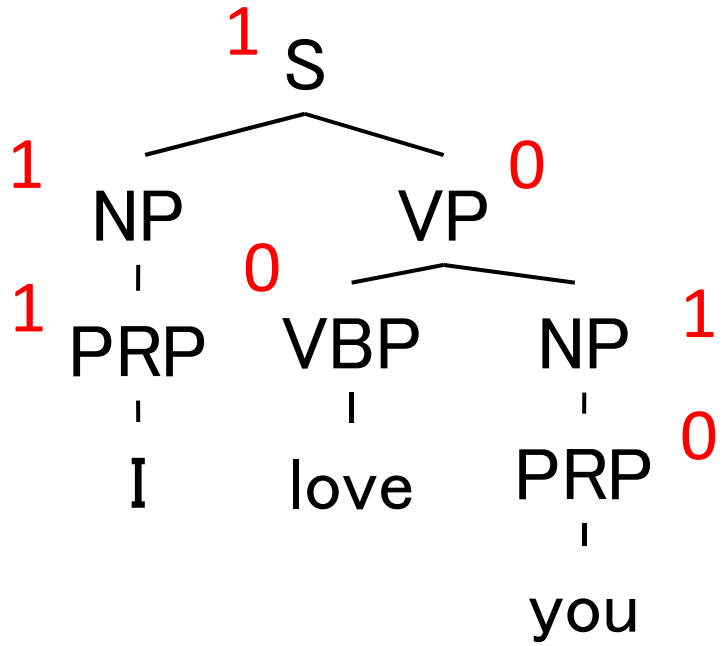
ギブスサンプリング



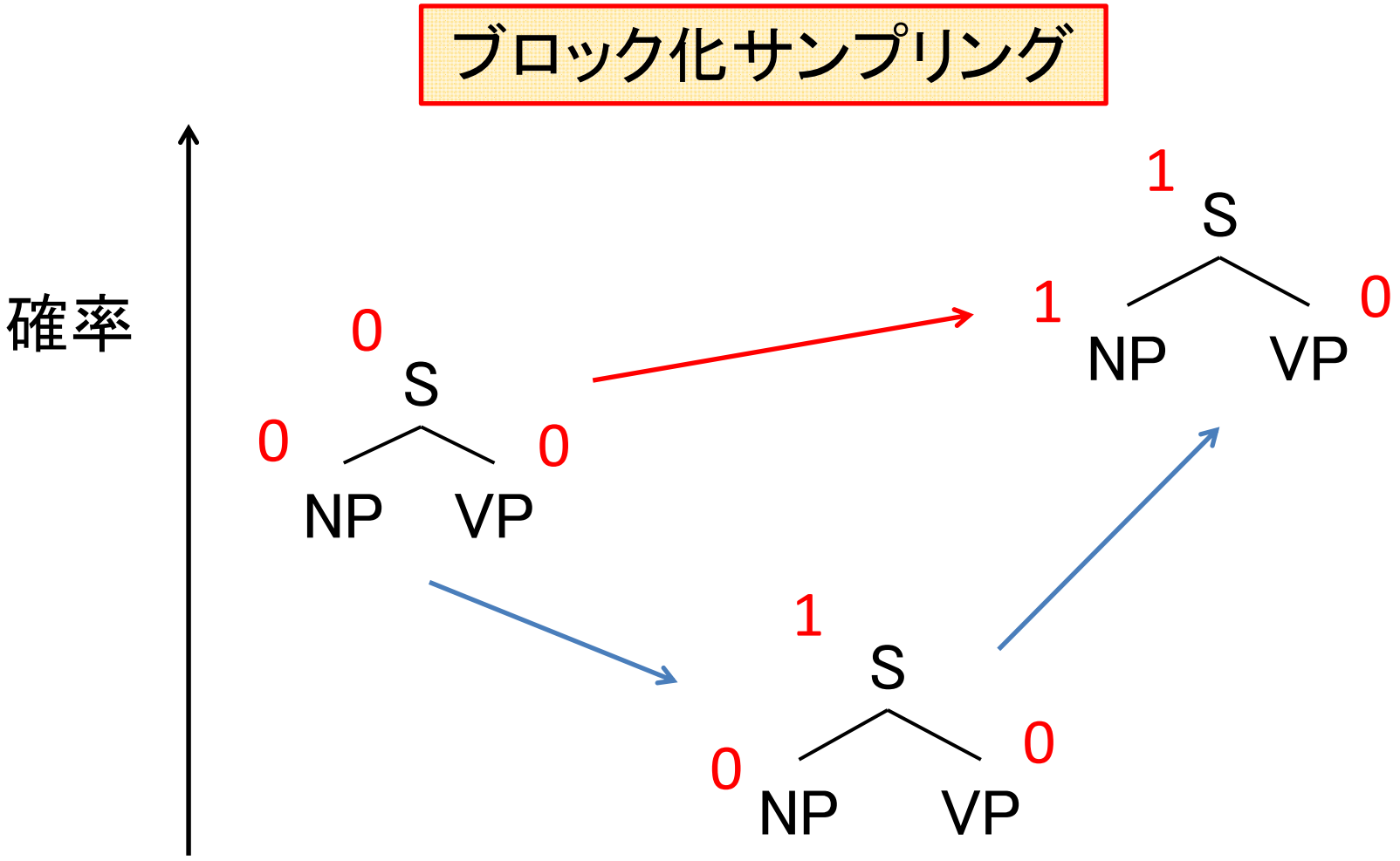
ギブスサンプリング



ギブスサンプリング



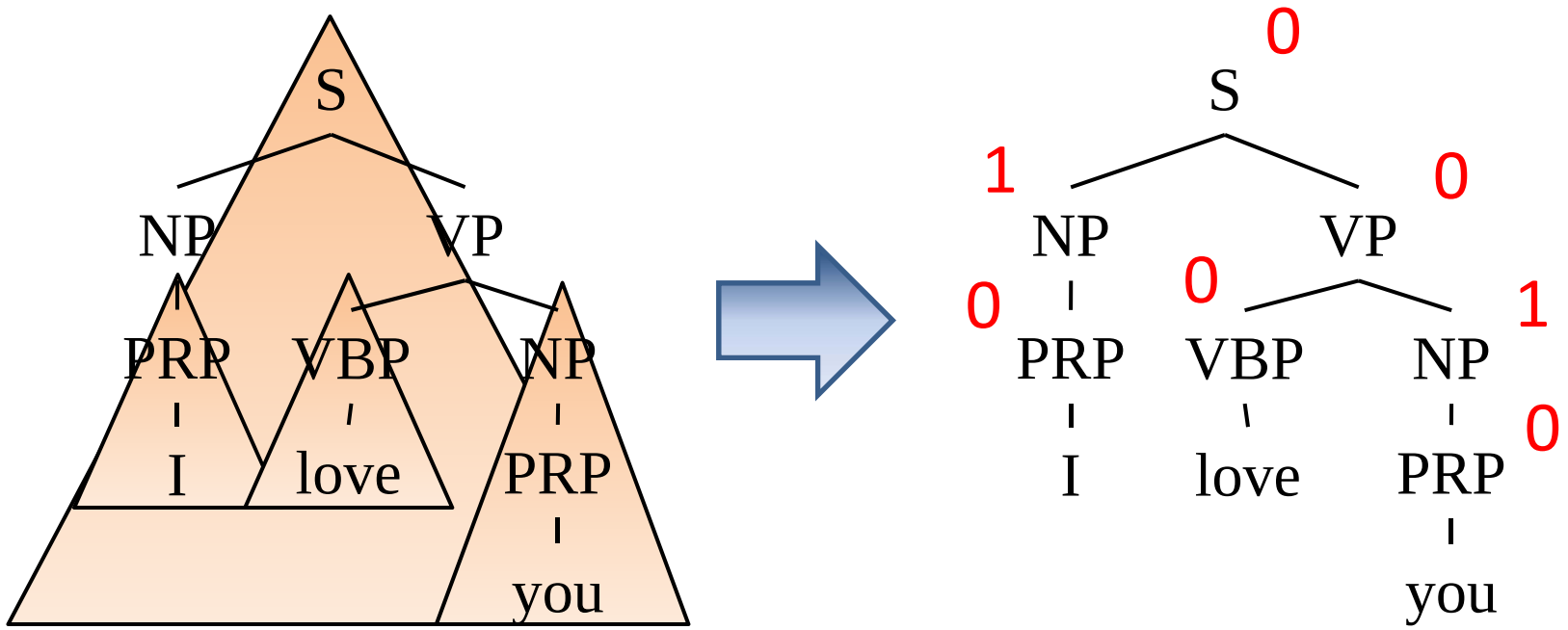
ギブスサンプリングは上手くいかない



ブロック化サンプリング

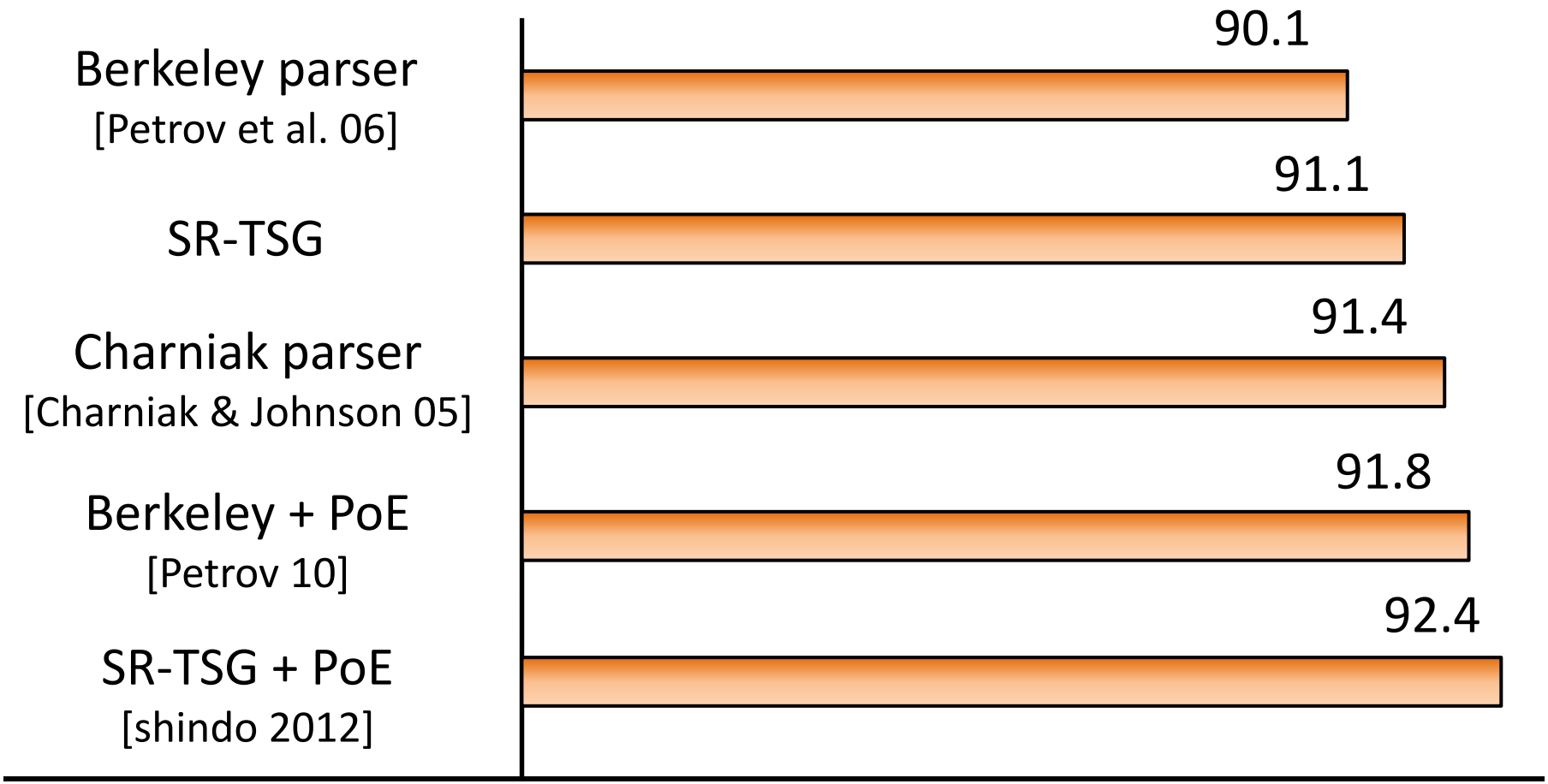
文の全ノードの潜在変数を一度に更新する
[Johnson 07]

動的計画法



Part4.　現在の到達点と今後の展開

現在の到達点



構文木コーパス： ４万文
言語： 英語

現在の到達点

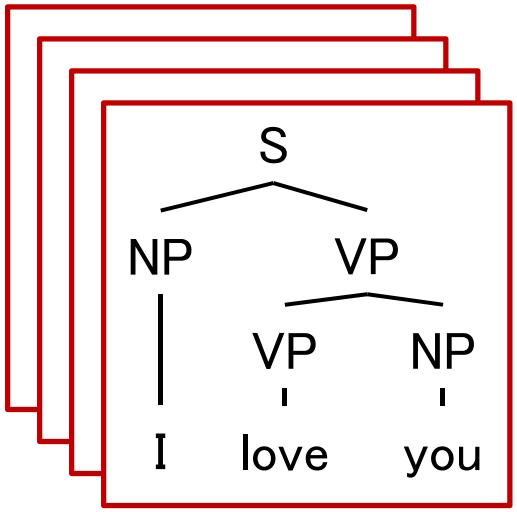
構文木コーパスが4万文あれば,
新聞記事のデータ(英語)に対して, 精度は90%を超える

・問題点:

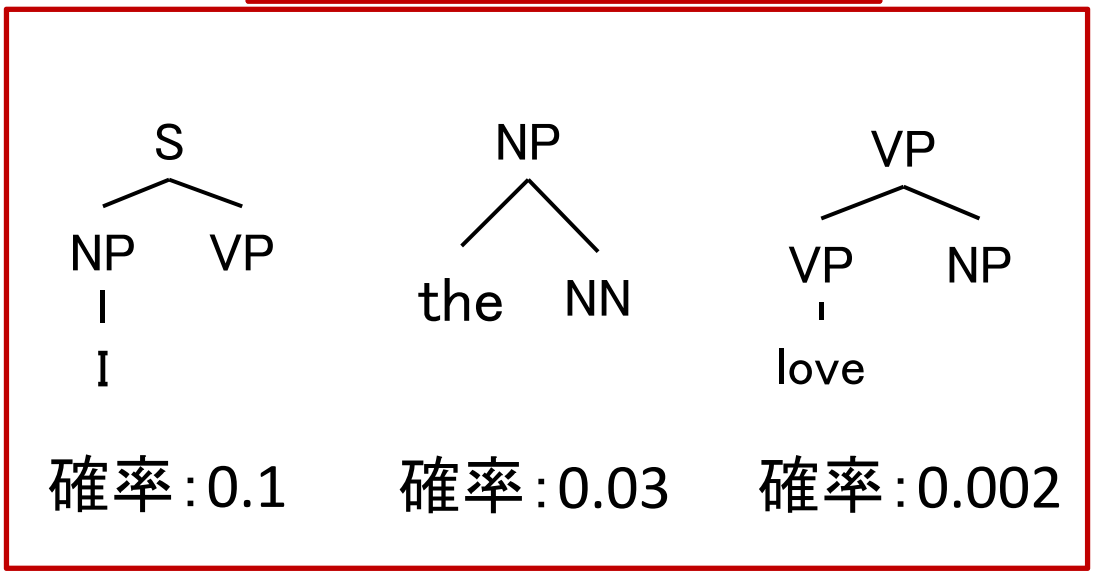
- ・新聞記事以外のデータは精度が低い
- ・Twitterなどの崩れた文は解析に失敗しやすい

教師あり学習から教師なし学習へ

構文木コーパス
（数万文）

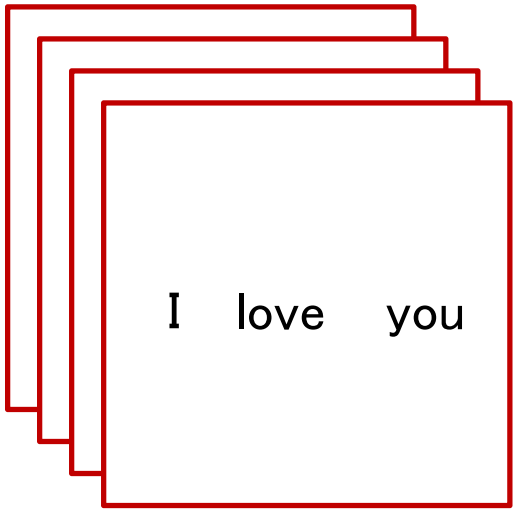


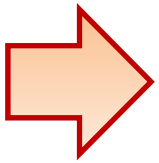
確率的文法モデル



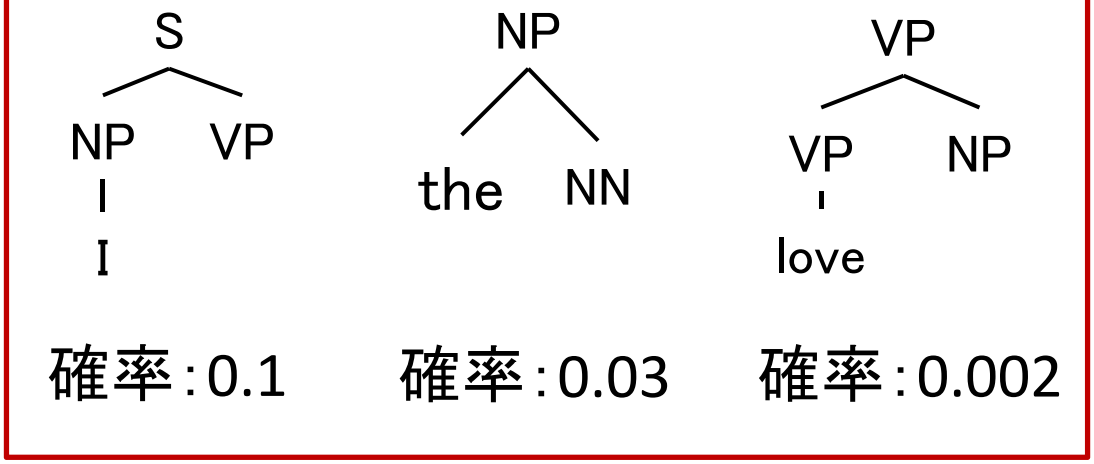
教師あり学習から教師なし学習へ

普通の文
（数億文以上）





確率的文法モデル



言語処理以外への展開

- 音楽データの解析
  - 木置換文法　[Bod 02]
  - 組合せ範疇文法　[Granroth-Wilding 12]

- バイオインフォマティクス（RNAの解析）
  - 文脈自由文法　[Eddy & Durbin 94]
  - 木接合文法　[Dowell 04]