シンボル細分化を適用した階層Pitman-Yor過程に基づく木置換文法獲得法と構文解析への応用

進藤 裕之[1]　宮尾 祐介[2]　藤野 昭典[1]　永田 昌明[1]

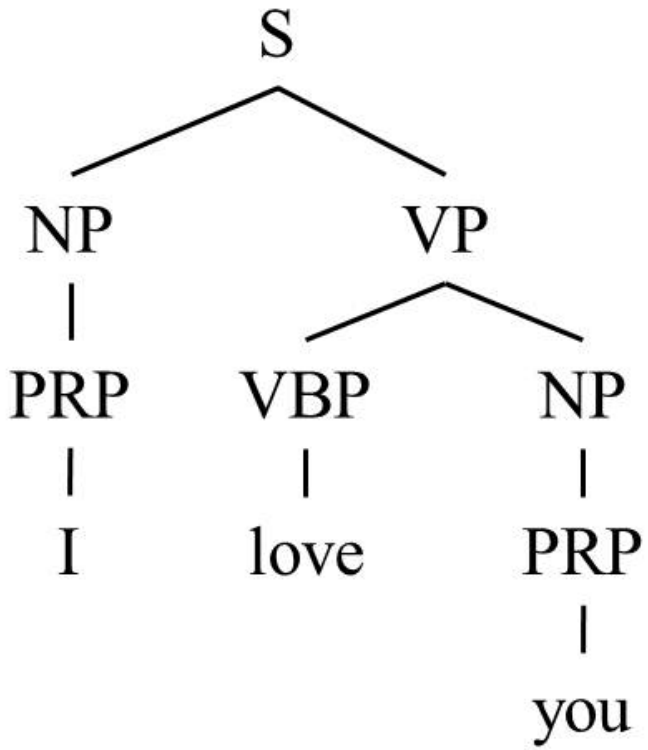
[1] NTT コミュニケーション科学基礎研究所
[2]国立情報学研究所

2012/03/16　言語処理学会第18回年次大会

目的

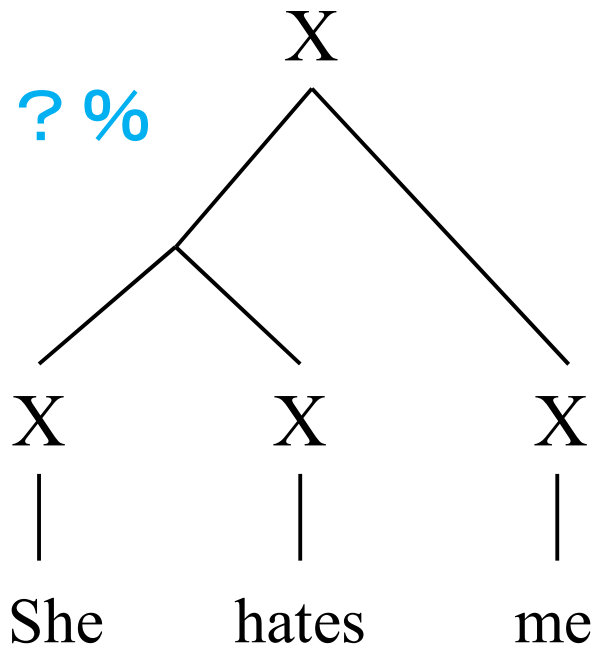
高精度な構文解析器(句構造)の実現

学習データ



テストデータ

解析精度:？%



目的

英語の構文解析器（state-of-the-art）

		精度
CFG + シンボル細分化	Collins Parser (1999)	88.5
	Berkeley Parser (2007)	90.1
	Charniak & Johnson's Parser (2005)	91.4
TSG + シンボル細分化	**提案手法： SR-TSG （最高精度）**	**92.4**

背景

・TSG（木置換文法）

・シンボル細分化

背景

- TSG（木置換文法）
- シンボル細分化

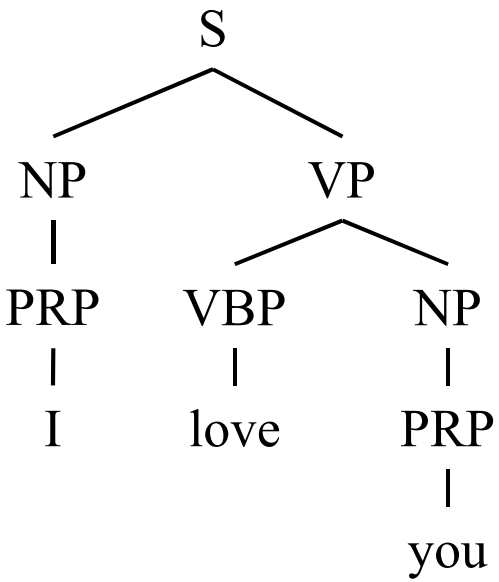
TSG（木置換文法）

Tree Substitution Grammars [Post+ 09, Cohn+ 09]

任意の大きさの部分木を基本単位とする構文木の生成モデル

TSG（木置換文法）

構文木



CFG（文脈自由文法）

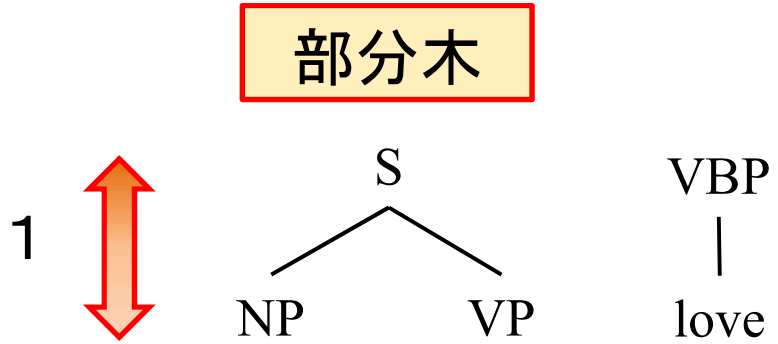
S → NP VP

VP → VBP NP

VBP → love

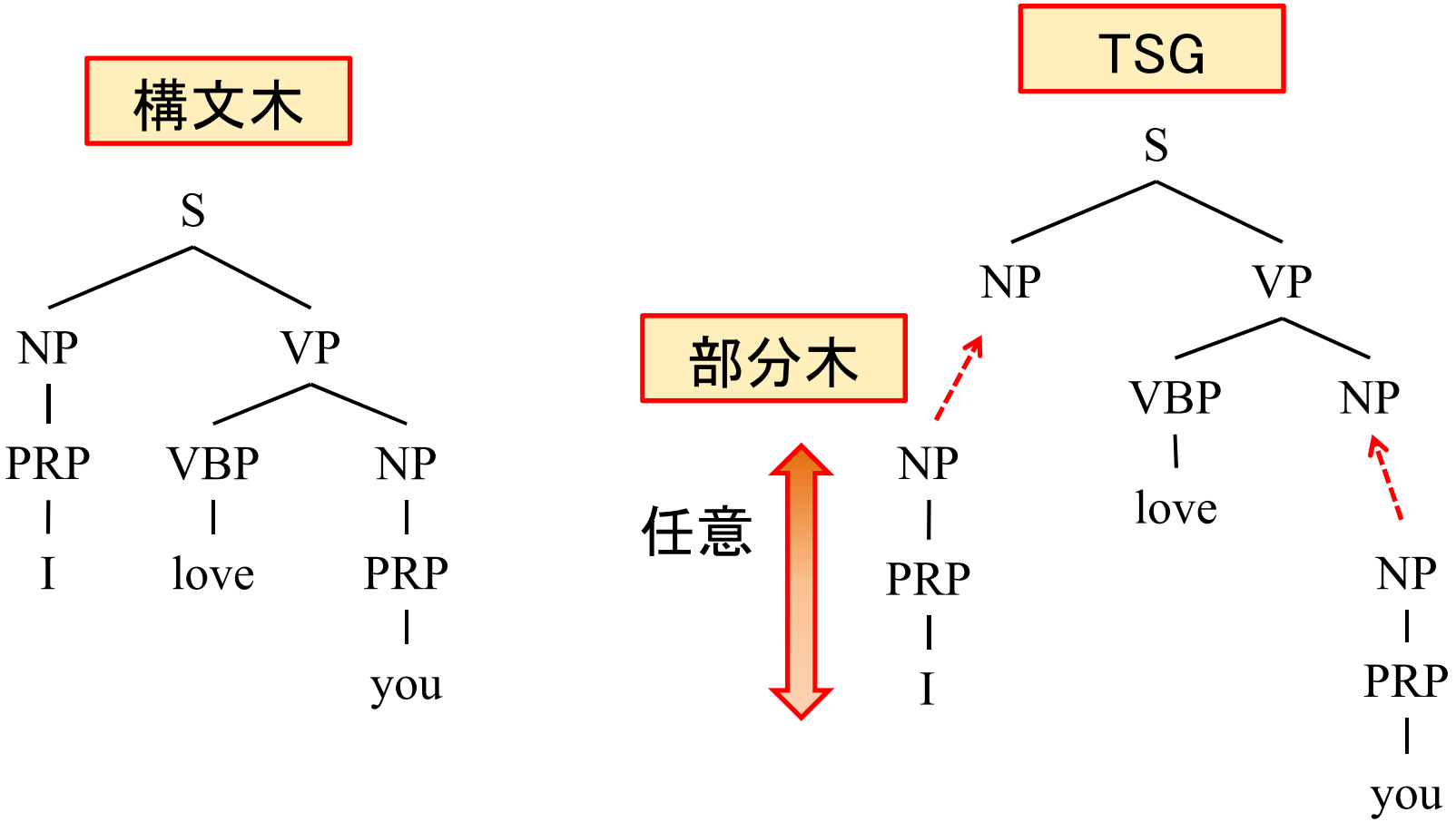
…

部分木



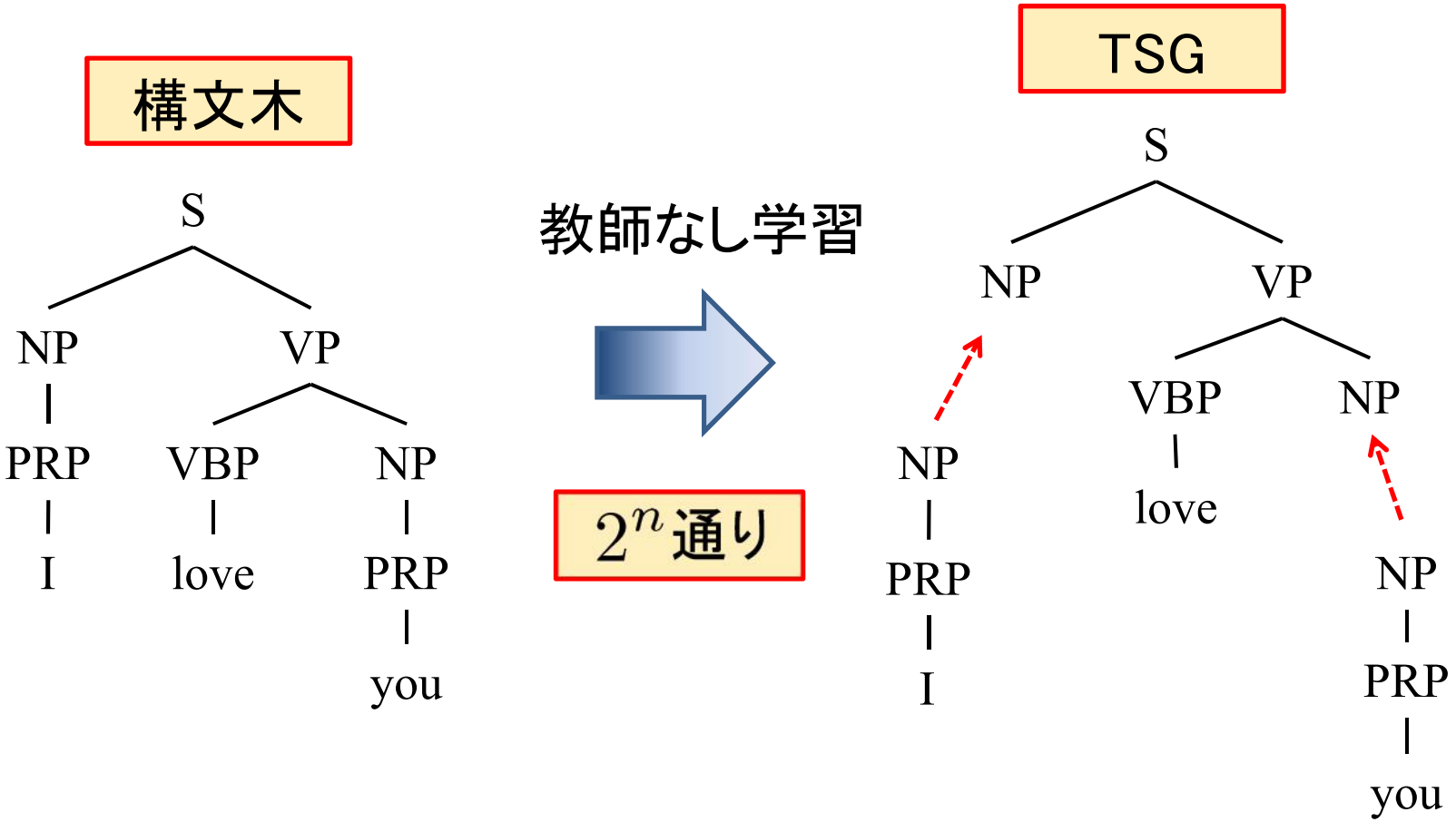
TSG（木置換文法）

TSG: CFG の拡張



TSG 部分木の学習

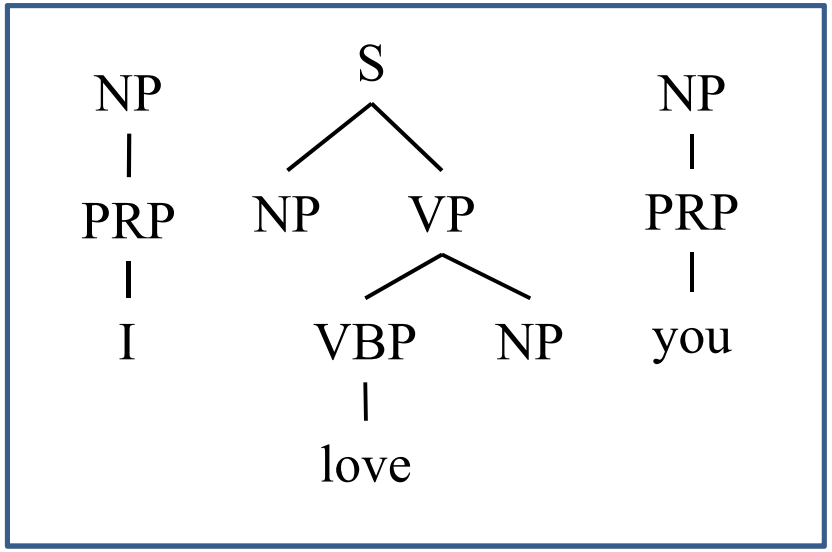
TSG 部分木の獲得　＝　教師なし構文木分割



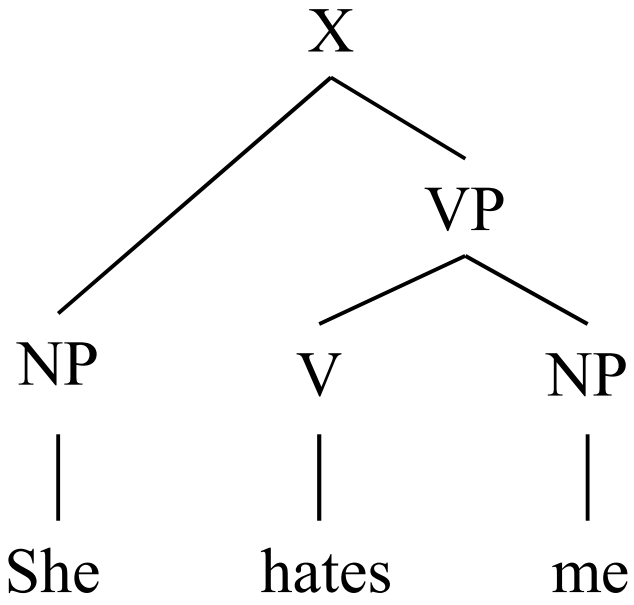
TSG に基づく構文解析

部分木を組み合わせて，確率の高い構文木を探索

獲得された部分木の集合



テストデータ



背景

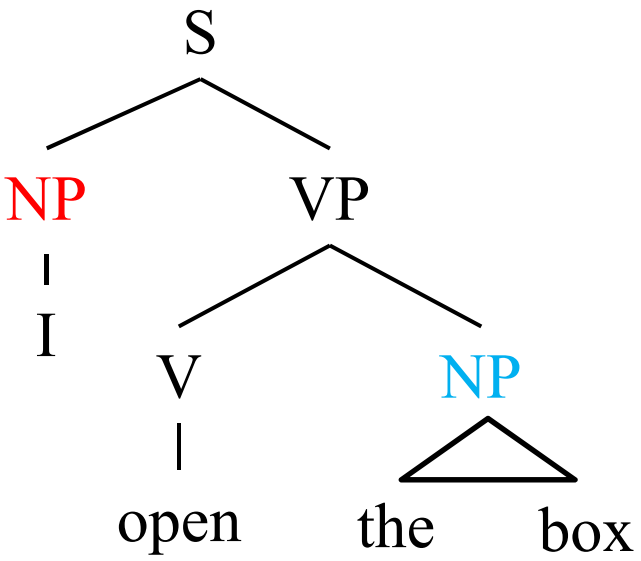
・TSG（木置換文法）

・シンボル細分化

シンボル細分化

Berkeley Parser, Charniak Parser の基盤となる手法

[Johnson 98, Collins 03, Matsuzaki+ 05, Petrov+ 06]



NP（主語）: I, he, she, ...

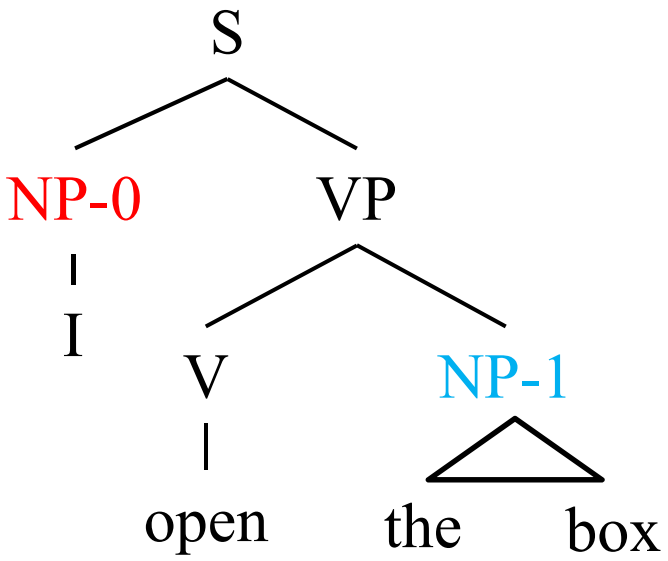
NP（目的語）: モノ

が生成されやすい

シンボル細分化

Berkeley Parser, Charniak Parser の基盤となる手法

[Johnson 98, Collins 03, Matsuzaki+ 05, Petrov+ 06]



NP−0（主語）:　I, he, she, ...

NP−1（目的語）: モノ

が生成されやすい

文脈情報をモデルに取り込む

構文解析の精度向上

既存研究の問題点

TSG ＋ シンボル細分化　⇒　データスパースネス

[Bansal+ 10] の研究

1.【前処理】 構文木コーパスの全シンボルを細分化

2. TSG 部分木の教師なし獲得

結果：☹ シンボル細分化の効果があまり表れない

（構文解析の精度があまり向上しない）

シンボル細分化によって，部分木の種類数が大幅に増大

限られた量の学習データでは，多くの部分木が一度も現れない

提案手法:

SR-TSG（シンボル細分化 TSG）

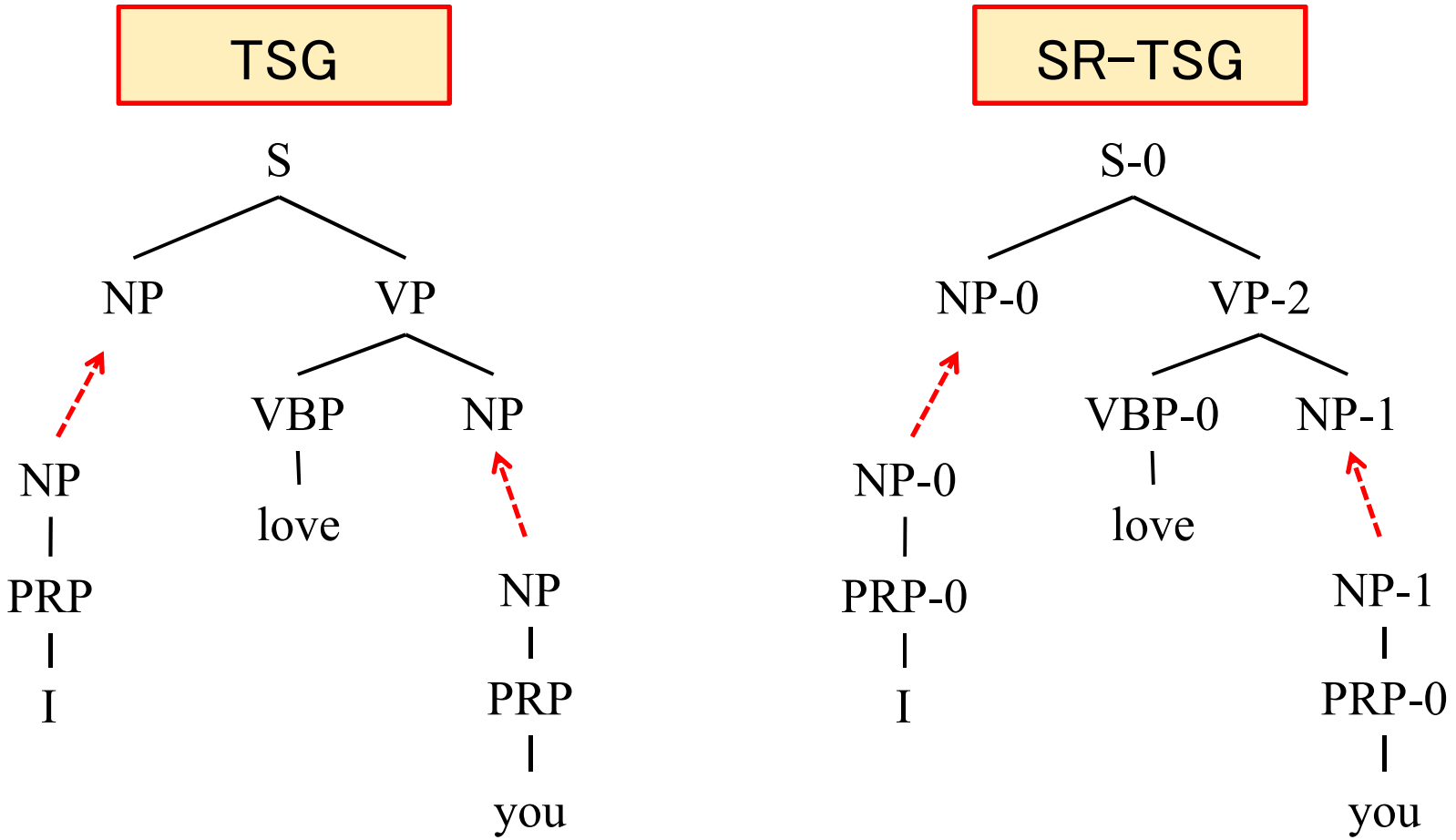
SR-TSG モデル

基本アイディアは3つ

1. シンボル細分化と TSG の統合モデル
2. 階層的なモデル構造を用いたスムージング
3. Pitman-Yor 過程によるデータに適応的な確率分布

SR-TSG モデル

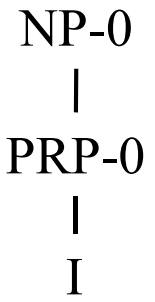
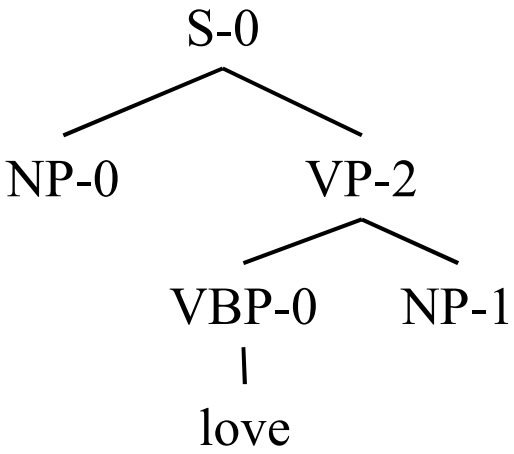
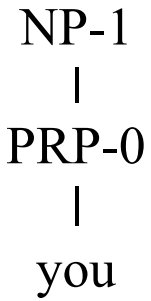
1. シンボル細分化と TSG の統合モデル



SR-TSG モデル

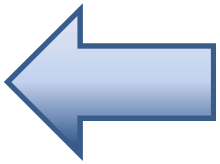
1. シンボル細分化と TSG の統合モデル

・シンボル細分化と TSG 部分木の同時確率

部分木			
確率	0.01	0.002	0.02

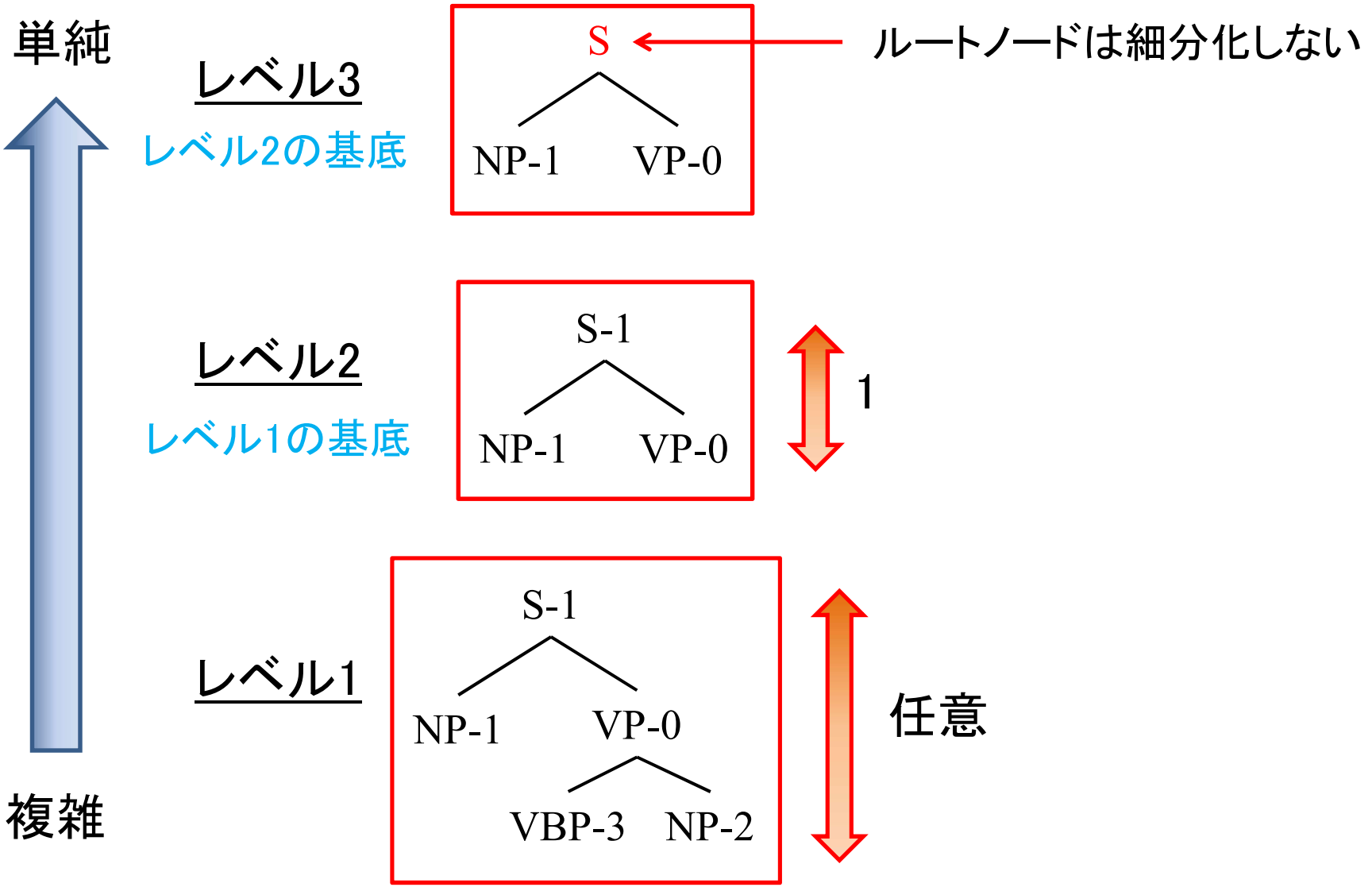
SR-TSG モデル

2. 階層的なモデル構造によるスムージング

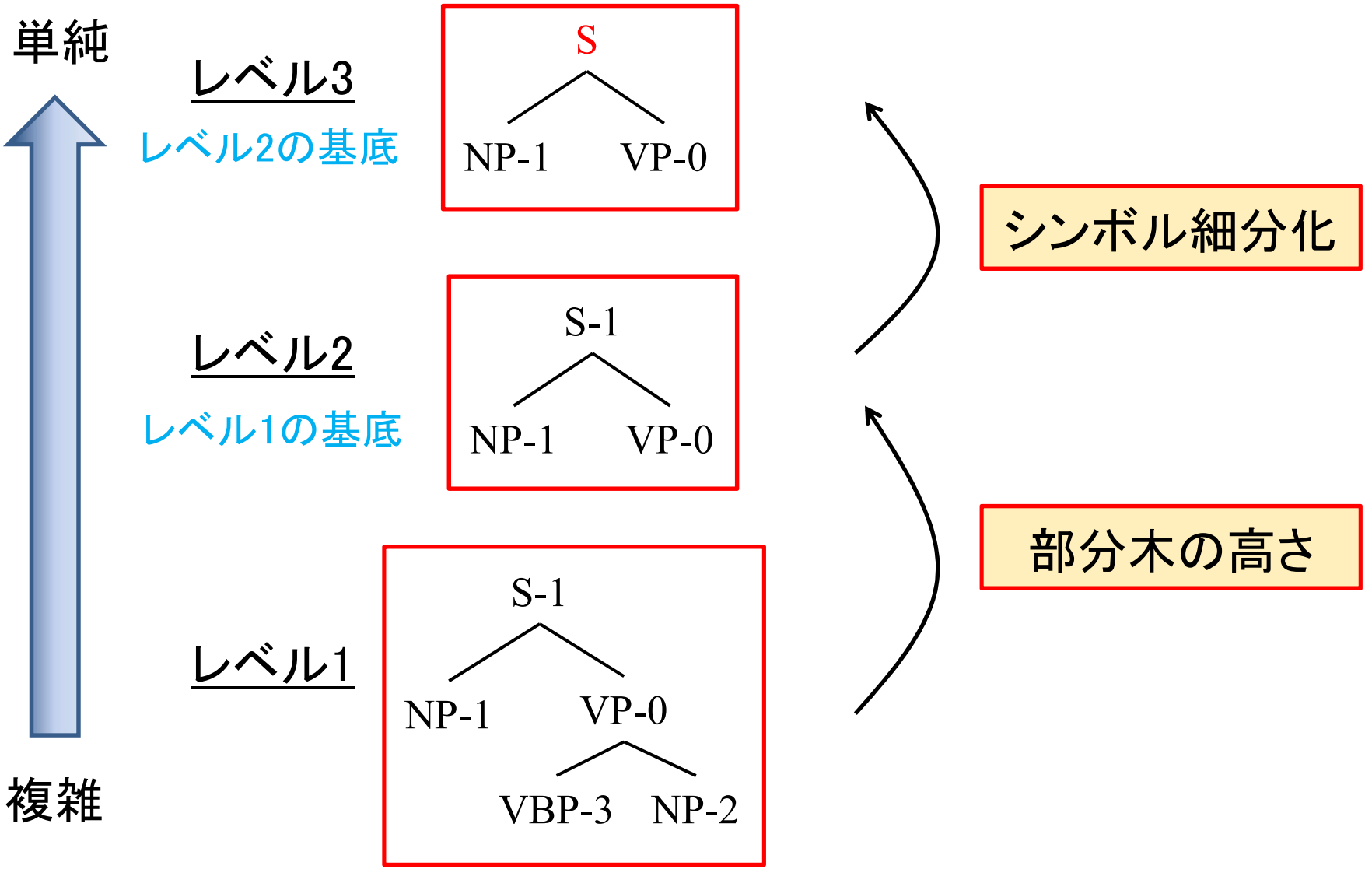
複雑な部分木の確率  単純な部分木の確率

スムージング

SR-TSG モデル

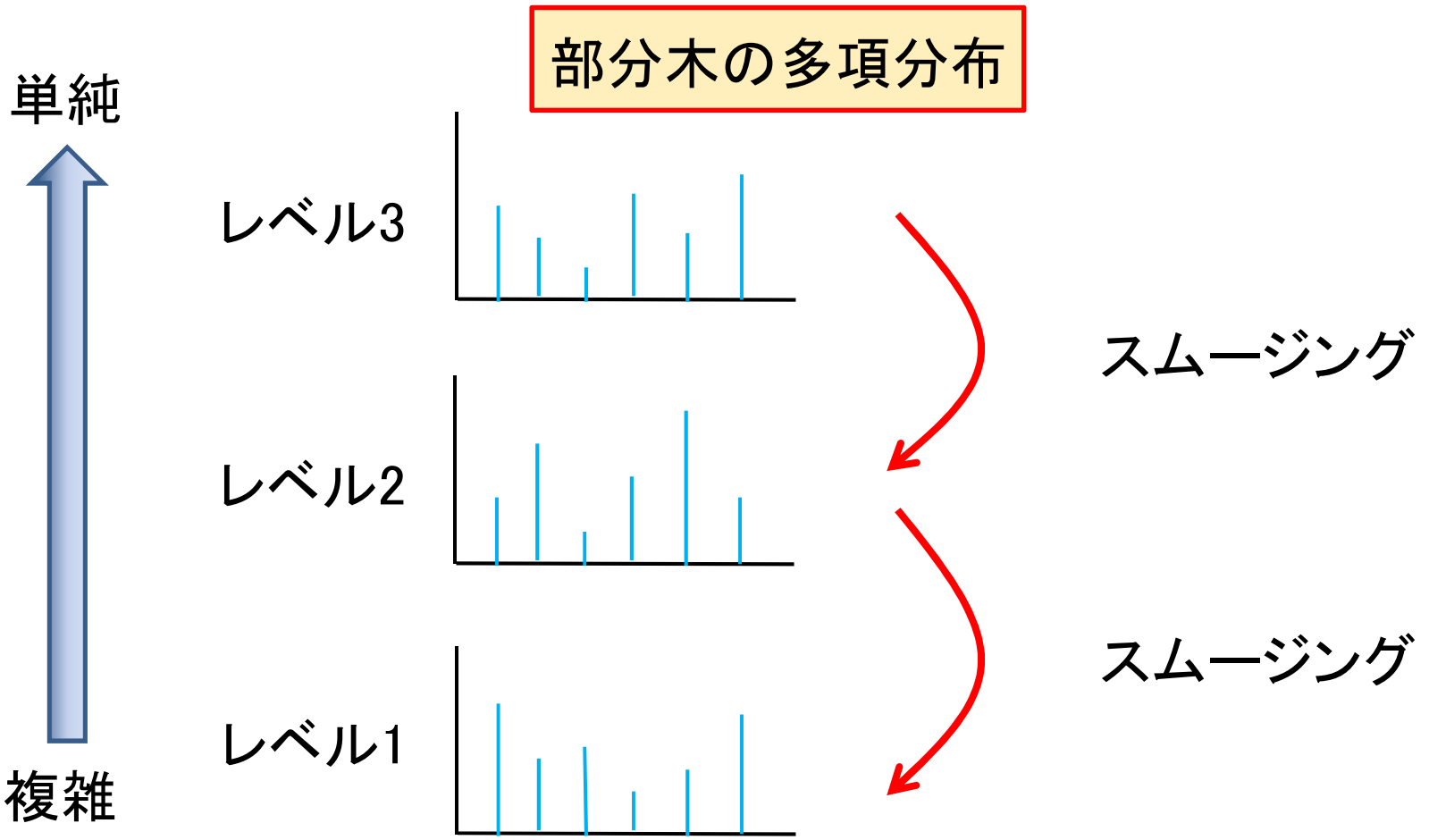

単純
複雑
レベル3
レベル2の基底
S
NP-1
VP-0
ルートノードは細分化しない
レベル2
レベル1の基底
S-1
NP-1
VP-0
1
レベル1
S-1
NP-1
VP-0
VBP-3
NP-2
任意

SR-TSG モデル


単純
複雑
レベル3
レベル2の基底
S
NP-1
VP-0
レベル2
レベル1の基底
S-1
NP-1
VP-0
レベル1
S-1
NP-1
VP-0
VBP-3
NP-2
シンボル細分化
部分木の高さ

SR-TSG モデル

☺ スムージングによりデータスパースネスを緩和

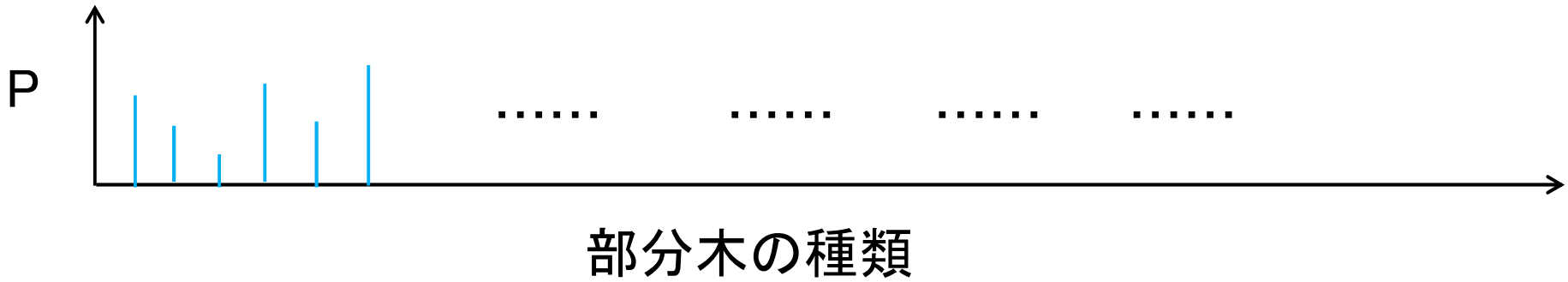


SR-TSG モデル

3. Pitman-Yor 過程によるデータに適応的な確率分布

・ノンパラメトリックベイズモデルの一種 [Pitman and Yor 97]

☹ 部分木の全可能性を列挙不可能



部分木の種類(次元数)が可変

データに応じて適切な部分木の種類数を推定可能

SR-TSG モデル

基本アイディアは3つ

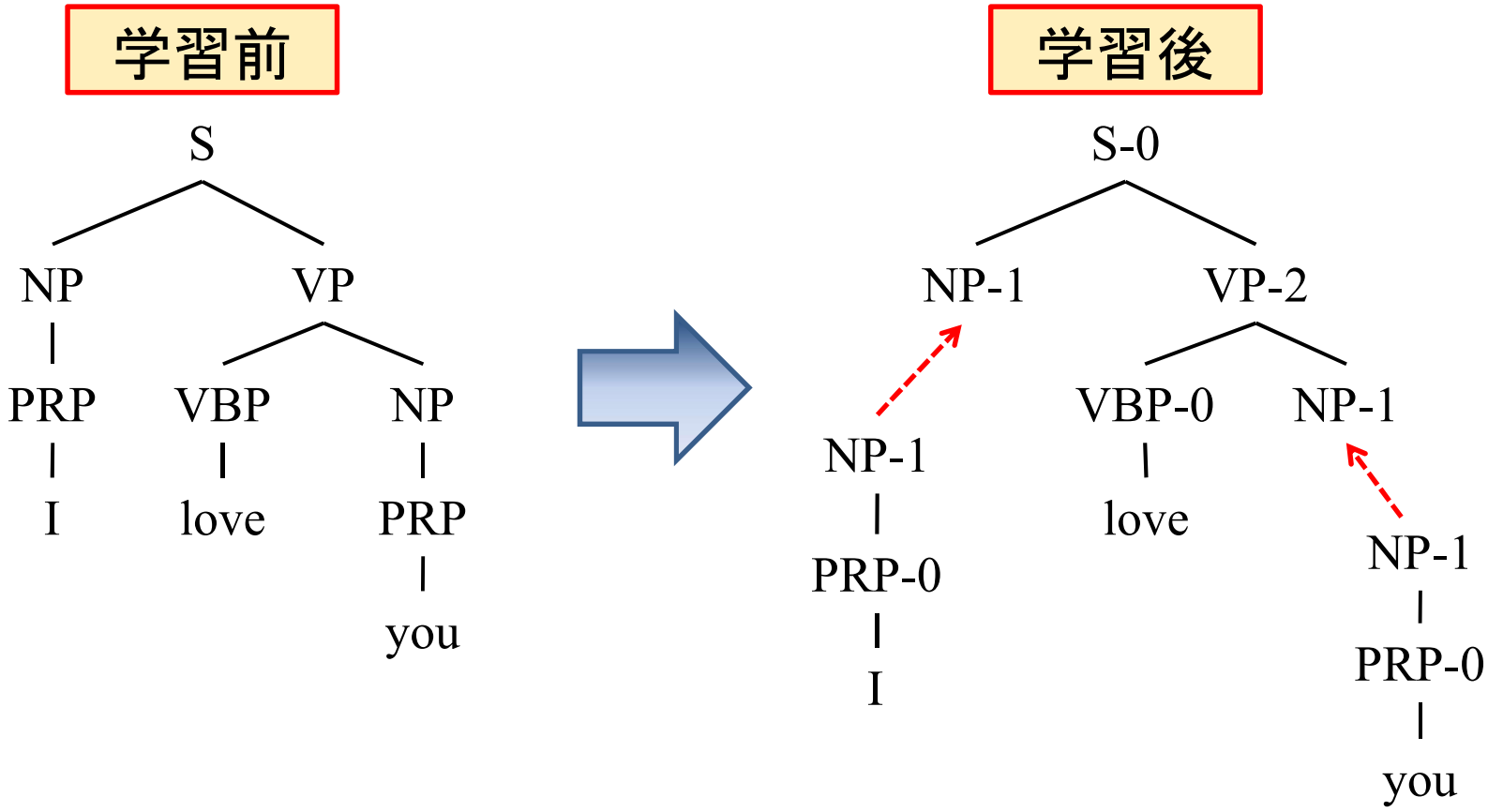
1. シンボル細分化と TSG の統合モデル
2. 階層モデルによるスムージング
3. Pitman-Yor 過程によるデータに適応的な確率分布

データスパースネスの緩和

部分木の種類数を適応的に決定

SR-TSG の学習

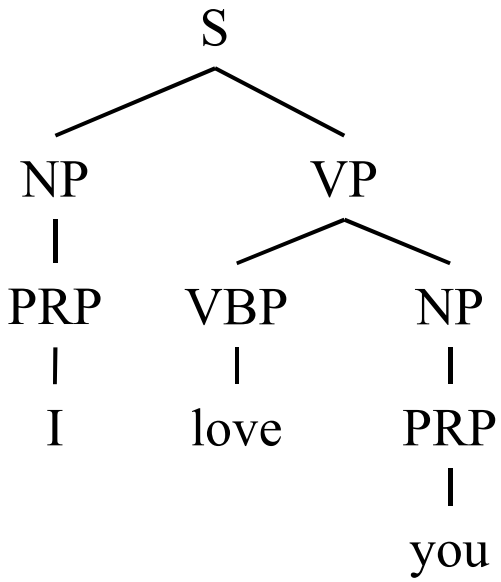
事後確率の最大化（MAP推定）



MCMC（マルコフ連鎖モンテカルロ法）

1. シンボル細分化の学習

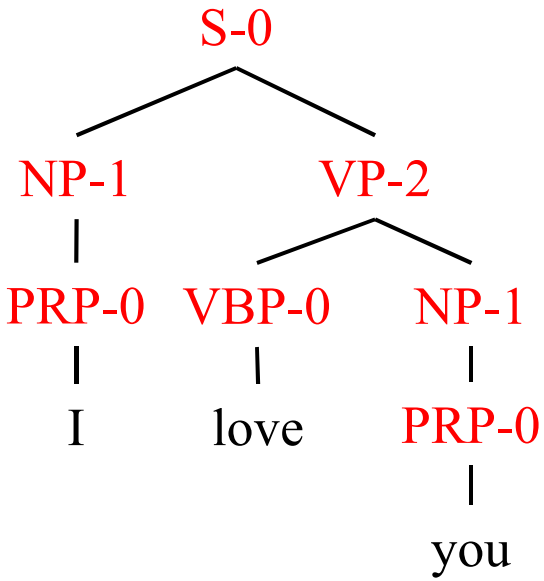
2. 構文木の分割による部分木の学習



MCMC（マルコフ連鎖モンテカルロ法）

1. シンボル細分化の学習

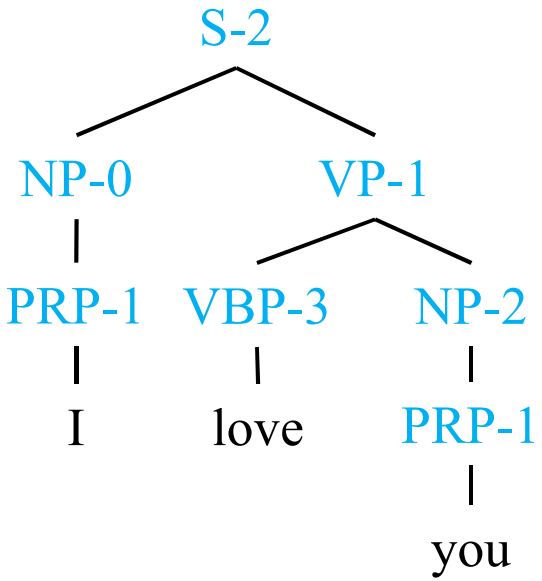
2. 構文木の分割による部分木の学習



MCMC（マルコフ連鎖モンテカルロ法）

1. シンボル細分化の学習

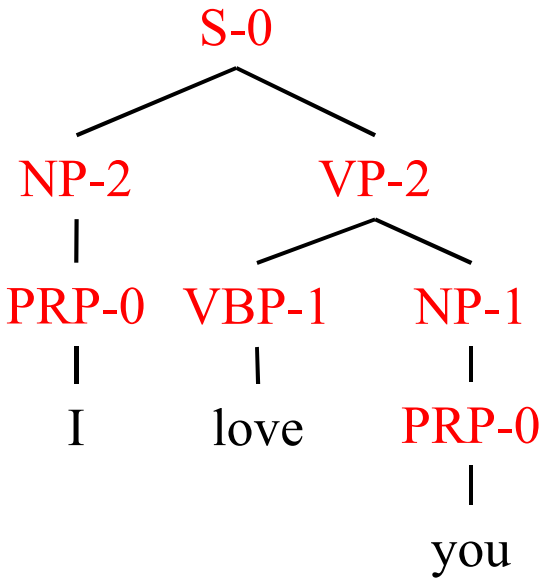
2. 構文木の分割による部分木の学習



MCMC（マルコフ連鎖モンテカルロ法）

1. シンボル細分化の学習

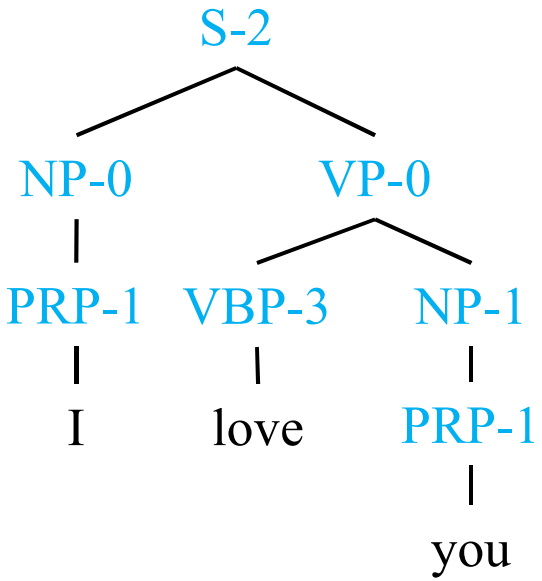
2. 構文木の分割による部分木の学習



MCMC（マルコフ連鎖モンテカルロ法）

1. シンボル細分化の学習

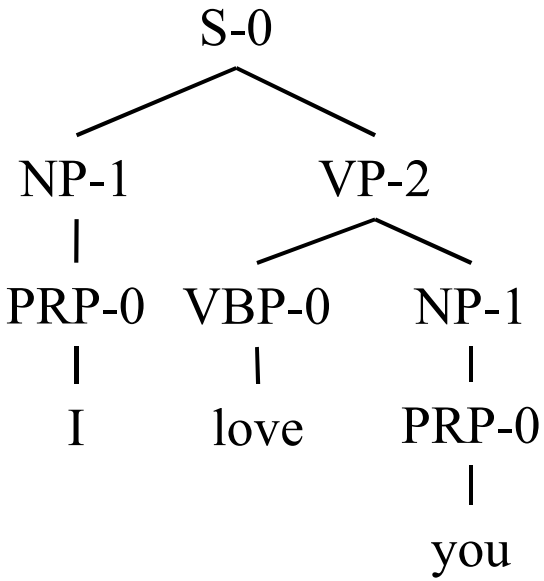
2. 構文木の分割による部分木の学習



MCMC（マルコフ連鎖モンテカルロ法）

1. シンボル細分化の学習

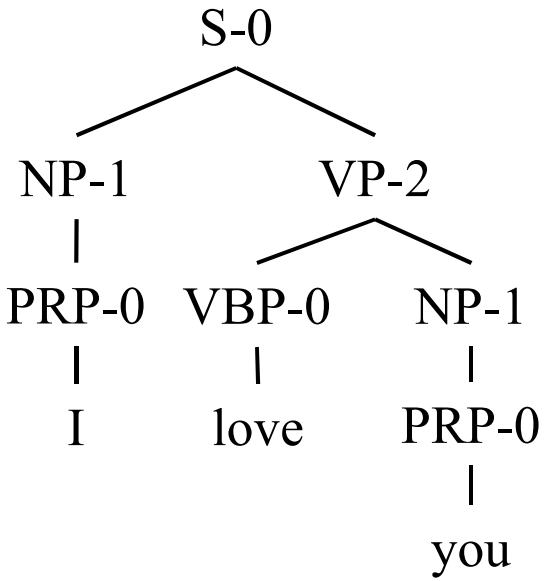
2. 構文木の分割による部分木の学習



MCMC（マルコフ連鎖モンテカルロ法）

1. シンボル細分化の学習

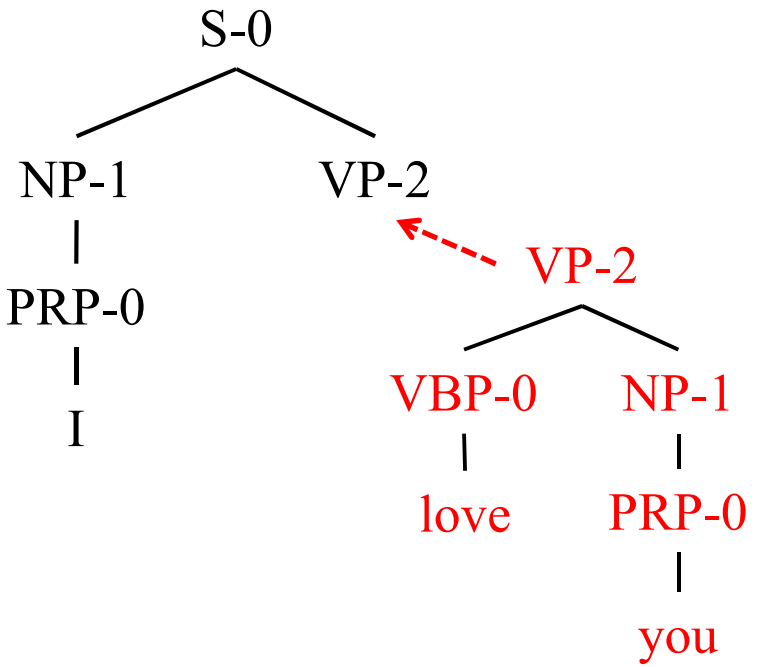
2. 構文木の分割による部分木の学習



MCMC（マルコフ連鎖モンテカルロ法）

1. シンボル細分化の学習

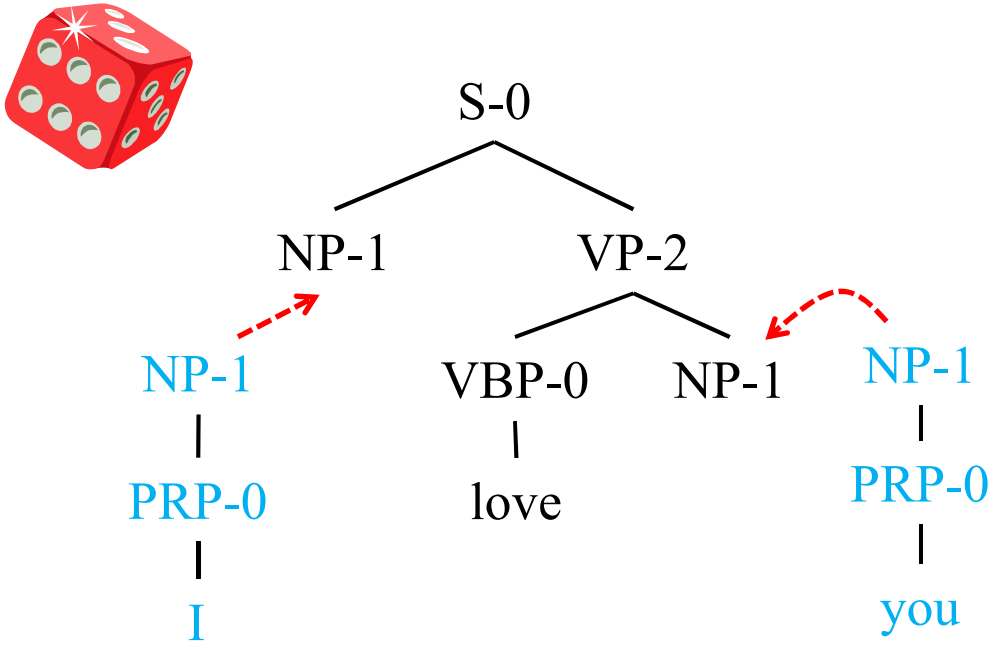
2. 構文木の分割による部分木の学習



MCMC（マルコフ連鎖モンテカルロ法）

1. シンボル細分化の学習

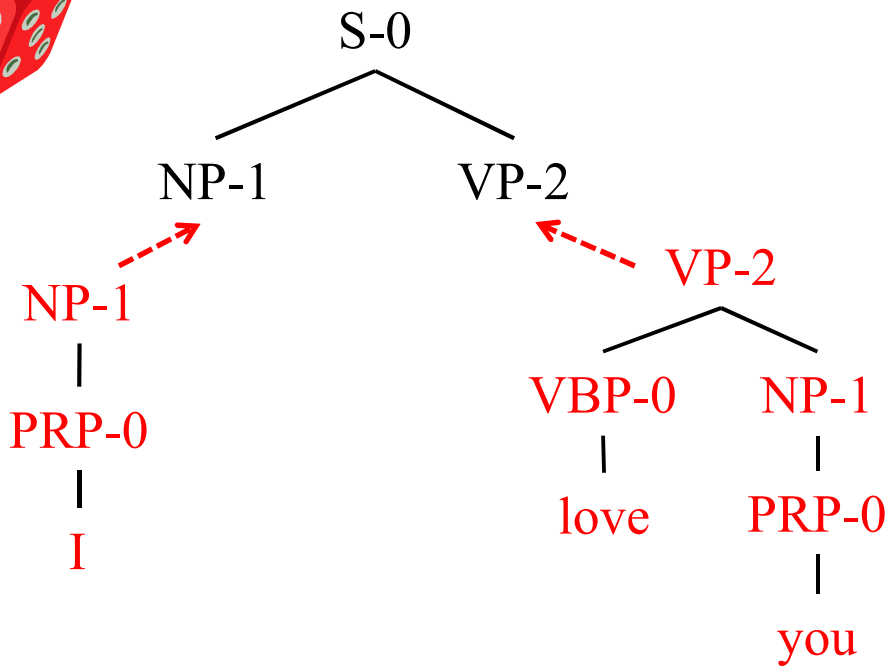
2. 構文木の分割による部分木の学習



MCMC（マルコフ連鎖モンテカルロ法）

1. シンボル細分化の学習

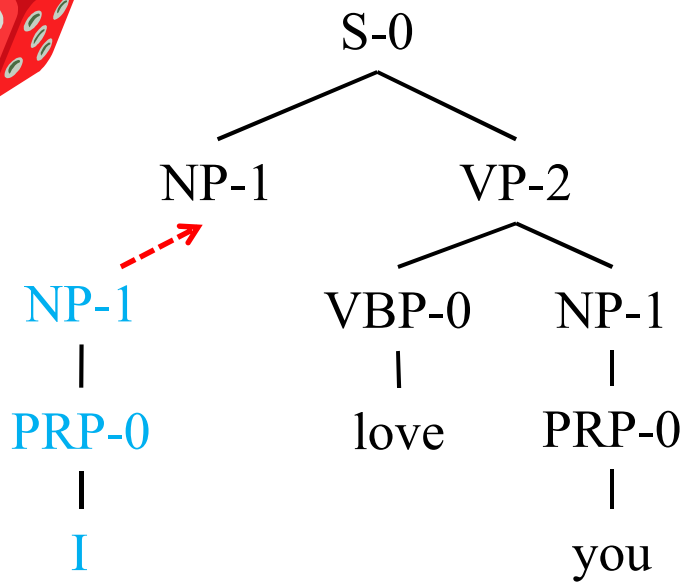
2. 構文木の分割による部分木の学習



MCMC（マルコフ連鎖モンテカルロ法）

1. シンボル細分化の学習

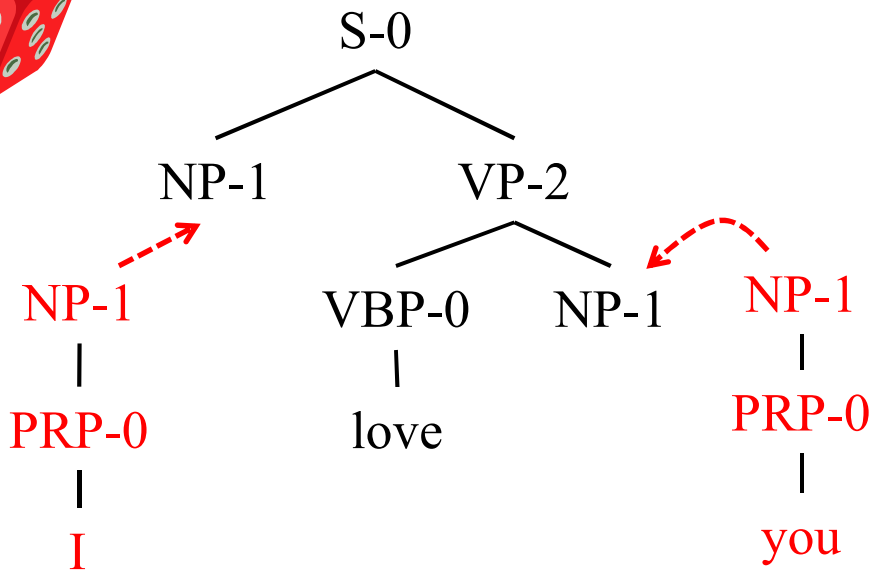
2. 構文木の分割による部分木の学習



MCMC（マルコフ連鎖モンテカルロ法）

1. シンボル細分化の学習

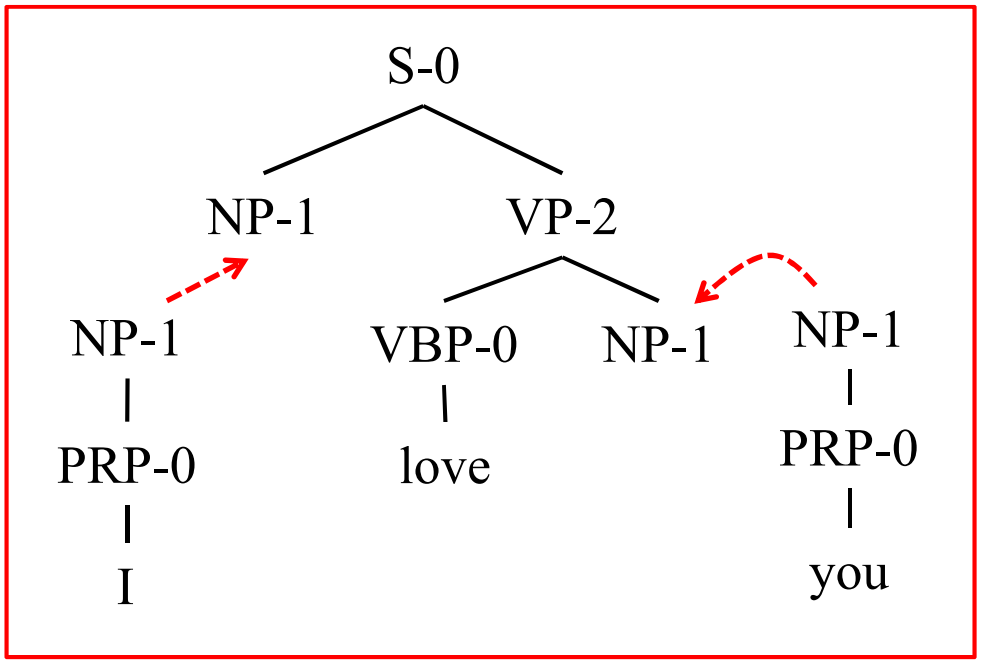
2. 構文木の分割による部分木の学習



MCMC（マルコフ連鎖モンテカルロ法）

1. シンボル細分化の学習

2. 構文木の分割による部分木の学習



構文解析実験

データセット

・WSJ Penn Treebank（英語の構文木コーパス）

- 学習データ　：約 40000 文
- テストデータ：約 2000 文

※ 最も標準的な英語の構文解析タスク

結果1

階層モデルの効果

	精度
SR-TSG（レベル1）	86.4
SR-TSG（レベル1＋2）	89.7
SR-TSG（レベル1＋2＋3）	91.1

・レベル1　　：スムージングなし

・レベル1＋2　　：１段階のスムージング

・レベル1＋2＋3　　：２段階のスムージング

結果２

シンボルが細分化された部分木の例

名詞

NNP-0	Corp.	Co.	Inc.
NNP-1	Brian	Howard	Christina
NNP-2	Feb.	Aug.	March

人の名前

動詞

VBZ-0	runs	comes	wins
VBZ-1	is	Is	gets
VBZ-2	says	means	claims

動詞＋that節

文法的に似た働きをする単語がクラスタ化されている

結果 3

代表的な構文解析手法との比較

	モデル	精度
TSG	Cohn et al. ’10	84.1
TSG + シンボル細分化	Bansal et al. ’10	88.1
CFG + シンボル細分化	Collins ’99	88.2
CFG + 識別モデル	Charniak & Johnson ’05	91.4
CFG + 識別モデル	Huang ’08	91.7
CFG + シンボル細分化	Petrov ’10	91.8
TSG + シンボル細分化	SR-TSG	92.4

SR-TSG は最高精度を達成

まとめ

TSG（木置換文法）に基づいた高精度な構文解析器

問題点：

・TSG + シンボル細分化　→　☹ データスパースネス

・膨大な可能性の部分木を計算機上で扱うことは困難

提案手法（SR-TSG）：

1. シンボル細分化と TSG の統合モデル
2. 階層的なスムージング
3. Pitman-Yor 過程に基づくデータに適応的な確率分布

結果：　☺ 英語の構文解析タスクで最高精度を実現