ALAGIN 機械翻訳セミナー
統語情報に基づく機械翻訳

Graham Neubig
奈良先端科学技術大学院大学 (NAIST)
2014 年 3 月 6 日

統語情報に基づく機械翻訳

- 今まで紹介した手法は構文解析を利用しない
- 構文解析は句を同定し、曖昧性を解消
  →訳の質向上につながると考えられる
- 原言語でも目的言語でも利用可能
- 主に２つの定式化手法
  - 同時文脈自由文法 (SCFG)
  - 木トランスデューサー

同時文脈自由文法

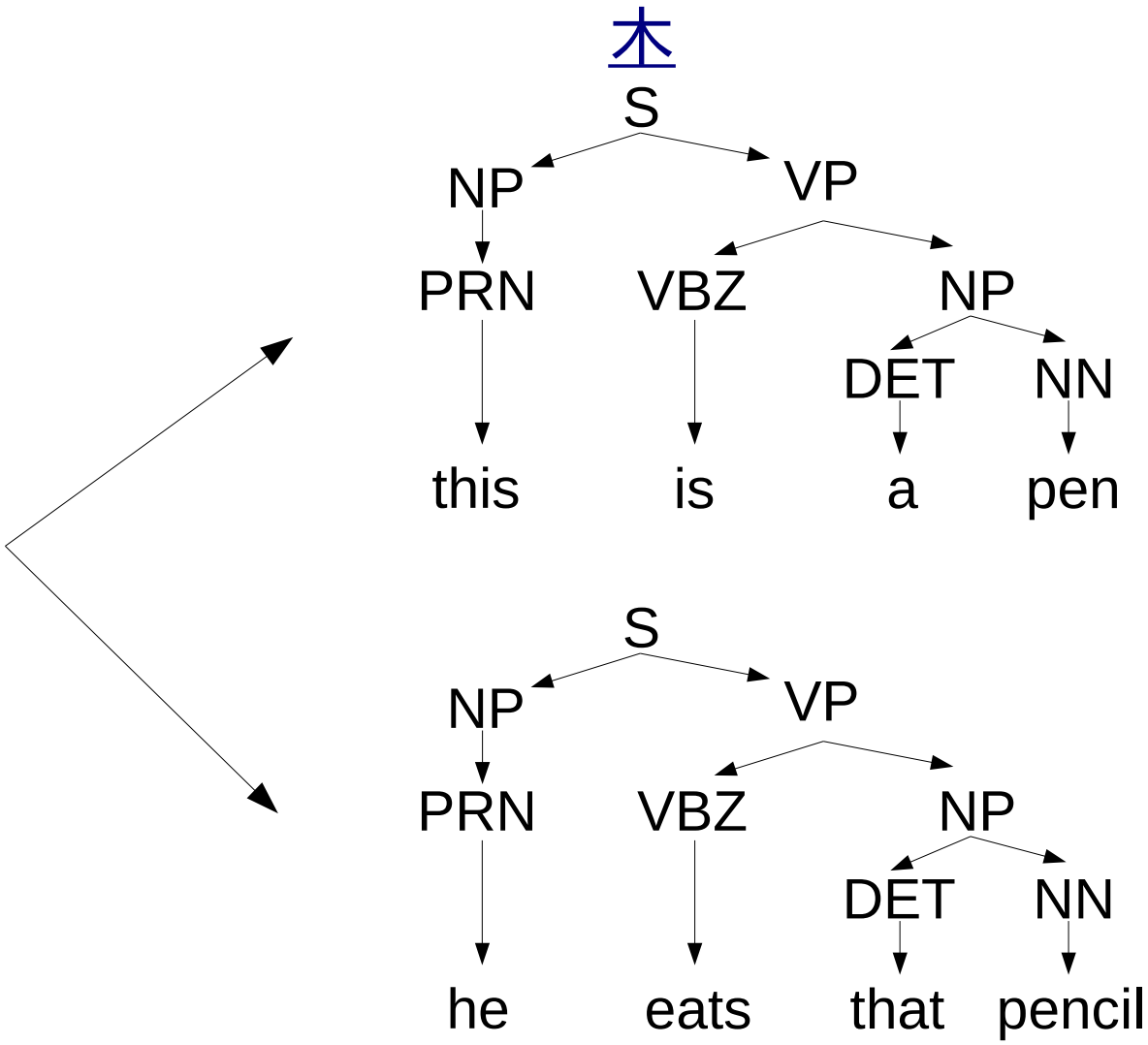
文脈自由文法 (CFG)

- 文を生成する規則を記述

文法

S → NP VP
NP → PRN
NP → DET NN
VP → VBZ NP
PRN → this
PRN → he
DET → a
DET → the
DET → that
NN → pen
NN → pencil
VBZ → is
VBZ → eats

木



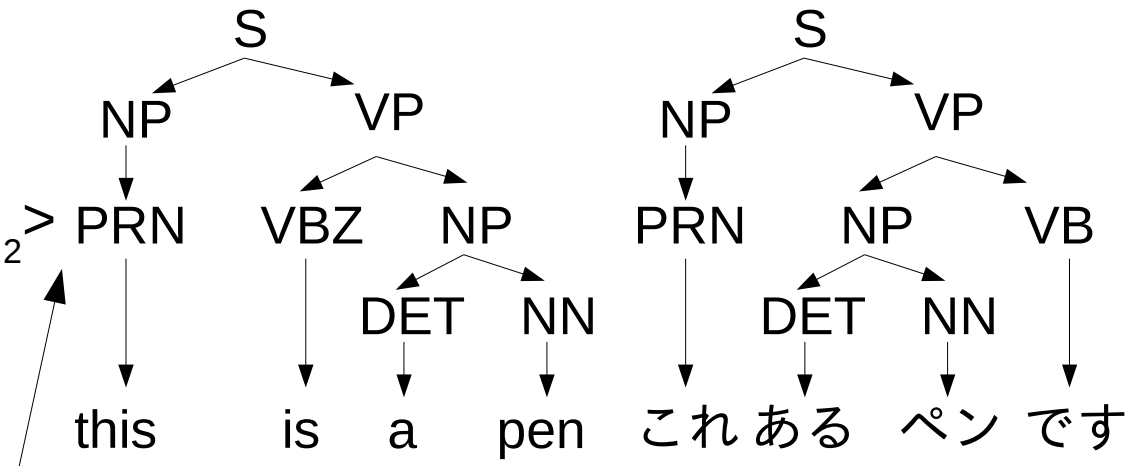
同時文脈自由文法 (SCFG)

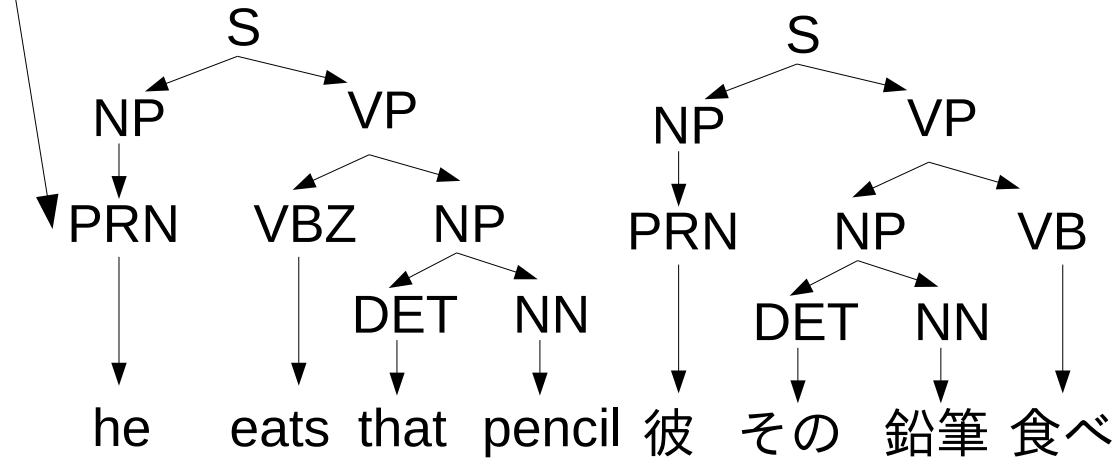
- 2 言語の文を同時に生成

文法

<S,S> → <NP$_1$ VP$_2$, NP$_1$ VP$_2$>
<NP,NP> → <PRN$_1$, PRN$_1$>
<NP,NP> → <DET$_1$ NN$_2$, DET$_1$ NN$_2$>
<VP,VP> → <VBZ$_1$ NP$_2$, NP$_2$ VB$_1$>
<PRN,PRN> → <this, これ>
<PRN,PRN> → <he, かれ>
<DET,DET> → <a, ある>
<DET,DET> → <the, その>
<DET,DET> → <that, その>
<NN,NN> → <pen, ペン>
<NN,NN> → <pencil, 鉛筆>
<VBZ,VB> → <is, です>
<VBZ,VB> → <eats, 食べます>

英語の木



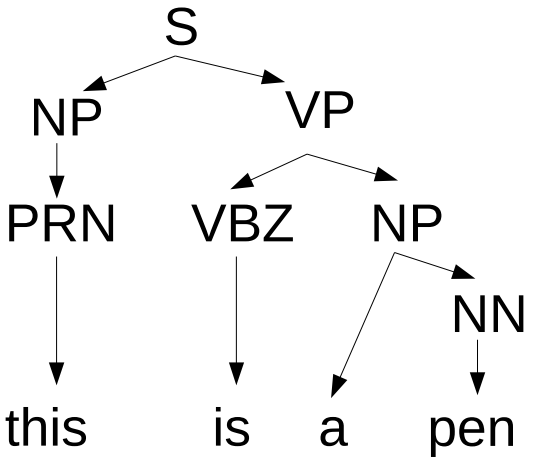


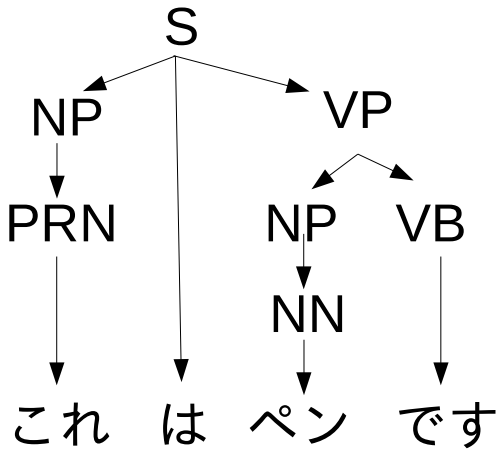
日本語の木

チョムスキー標準形以外の SCFG

- チョムスキー標準形で A → B C や A → x で限定
- チョムスキー標準形だけでは表しにくい言語現象も

<S,S> → <NP_1 VP_2, NP_1 は VP_2>
<NP,NP> → <the NN_1, NN_1>
<VP,VP> → <VBZ_1 NP_2, NP_2 を VB_1>
<NN,NN> → <a pen, ペン >



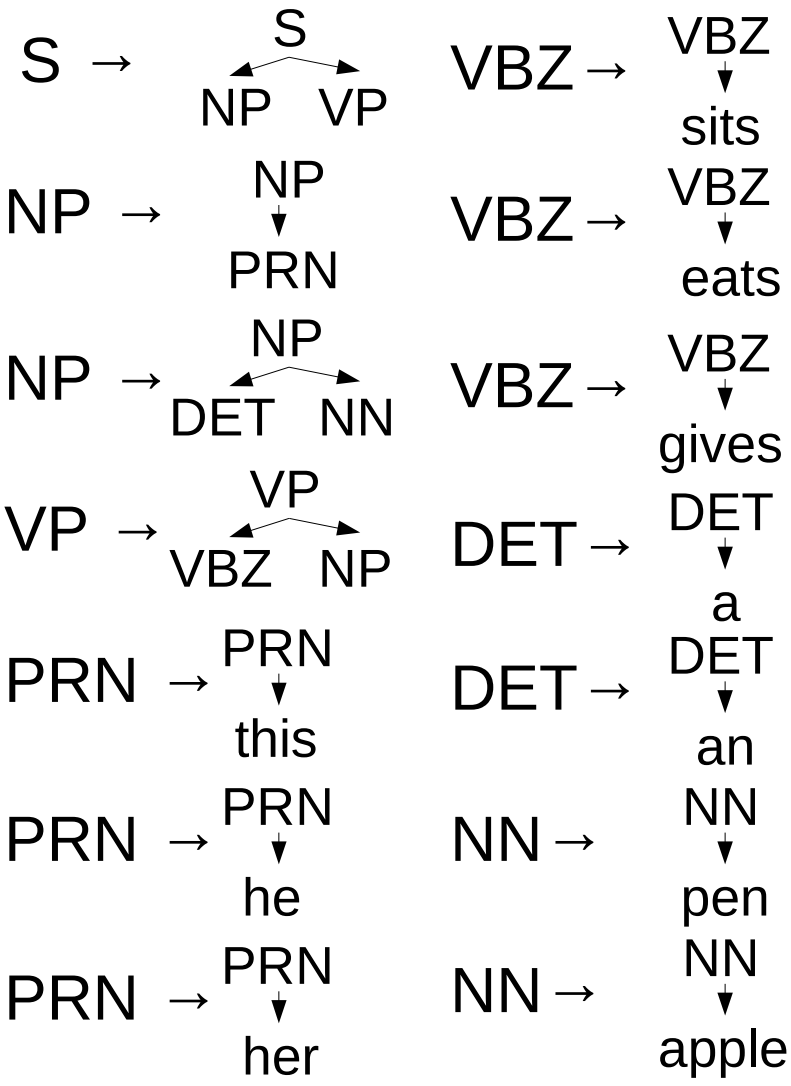


同期木置換文法

木置換文法 (TSG)

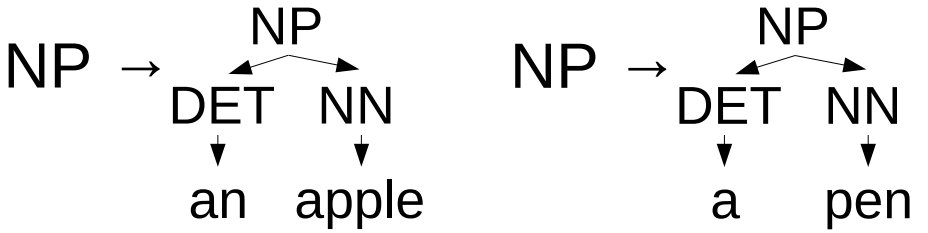
- 部分木を基本単位とした文法

CFG に含まれるルール

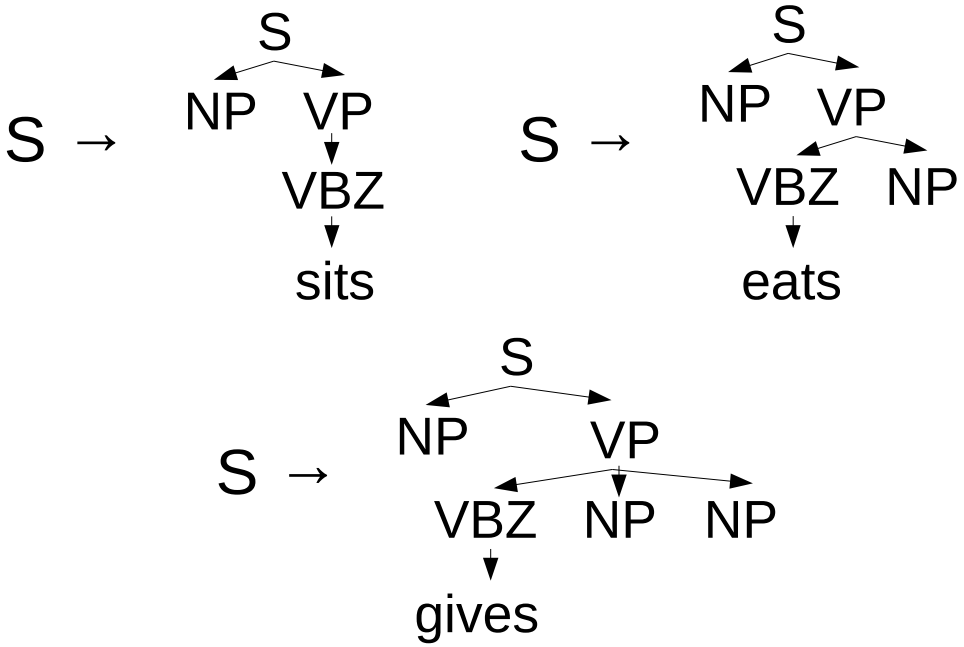


CFG に含まれないルール

冠詞の一致



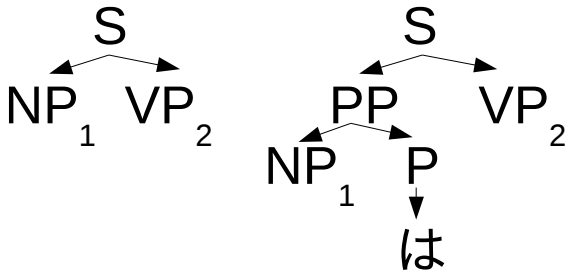
自動詞・他動詞等



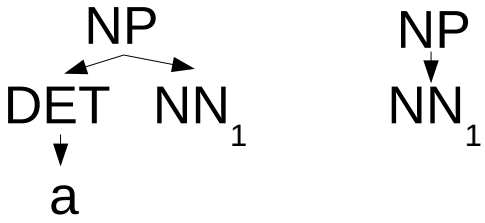
同期木置換文法 (STSG)

- 2 言語に渡る木置換文法

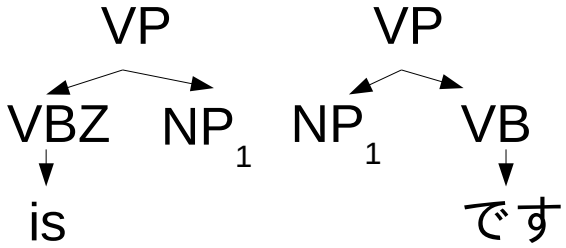
<S,S> →



<NP,NP> →

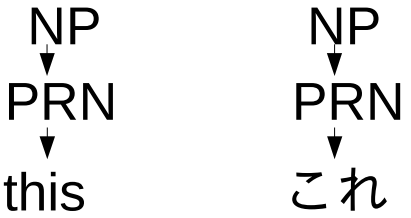


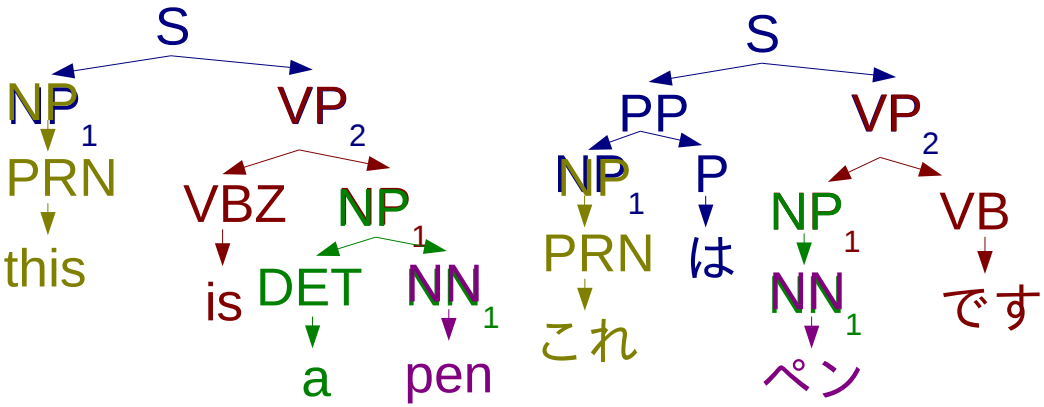
<VP,VP>→



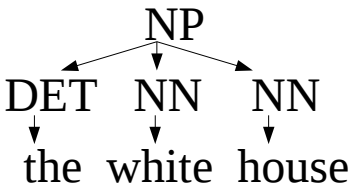
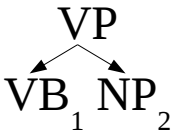
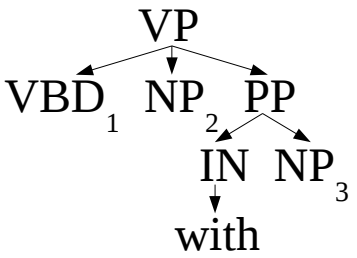
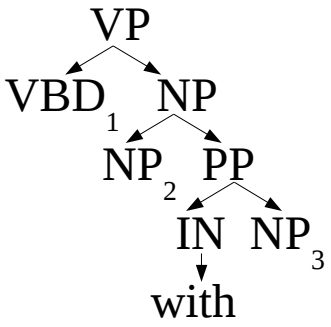
<NN,NN>→

<NP,NP>→





SCFG vs. STSG

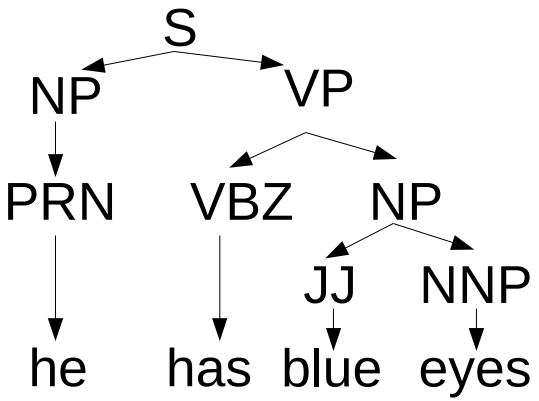
SCFG	STSG	目的言語
NP → the white house		ホワイト ハウス
VP → VB_1 NP_2		X_2 を X_1
VP → VBD_1 NP_2 with NP_3		X_3 で X_2 を X_1
VP → VBD_1 NP_2 with NP_3		X_3 の ある X_2 を X_1

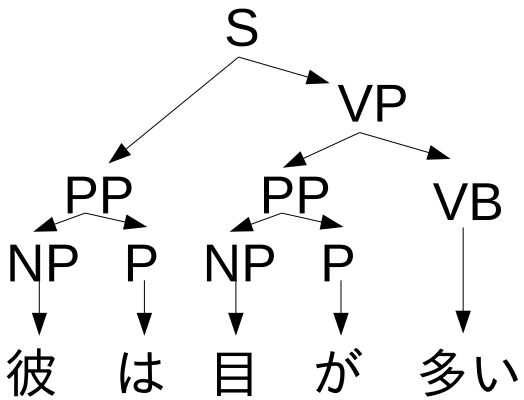
文法の不一致と対策

文法の不一致

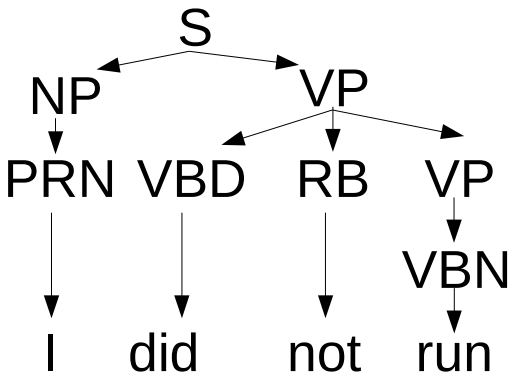
- 文法が合わない場合が多い

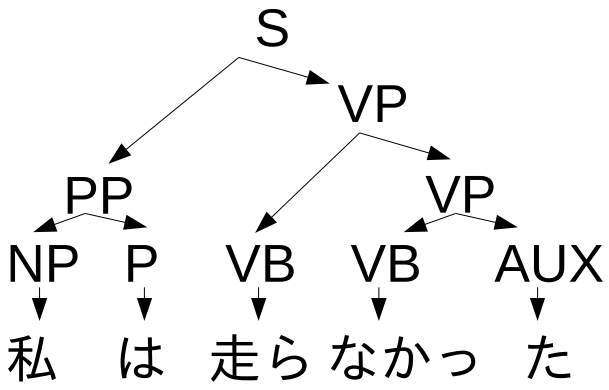
主辞の交代





構造の差





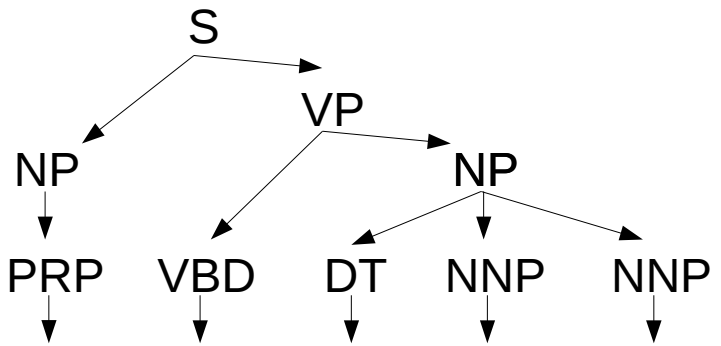
構文解析誤り

翻訳モデルの種類

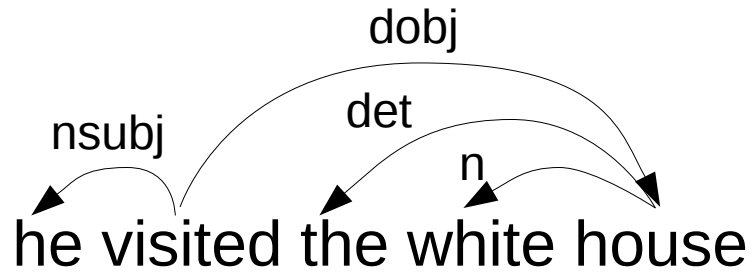
string

he visited the white house

tree



dependency

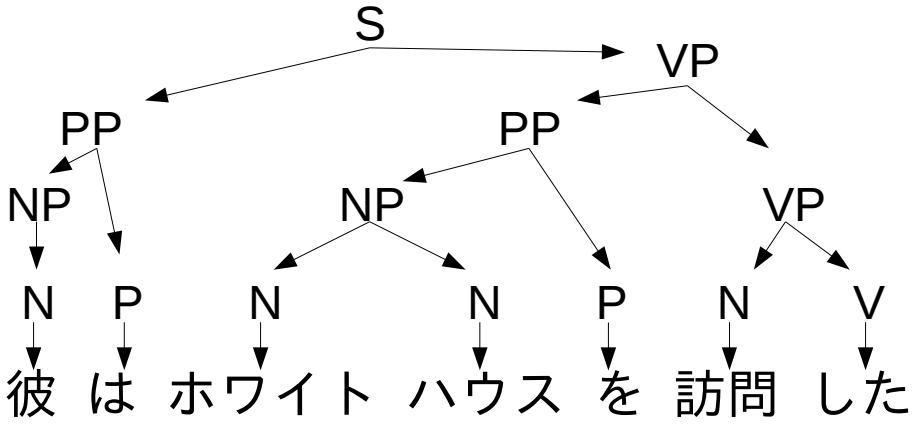


to

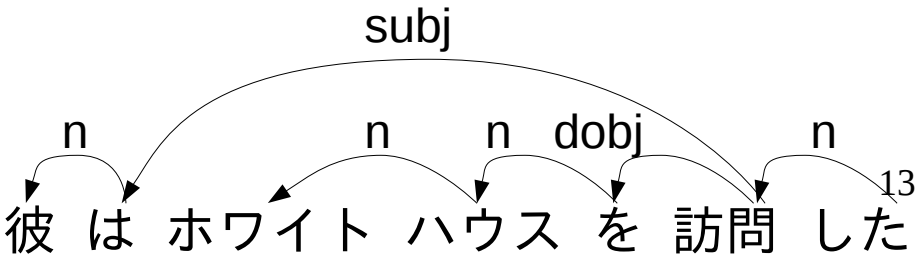
string

彼 は ホワイト ハウス を 訪問 した

tree



dependency



目的言語の構文情報を利用する翻訳

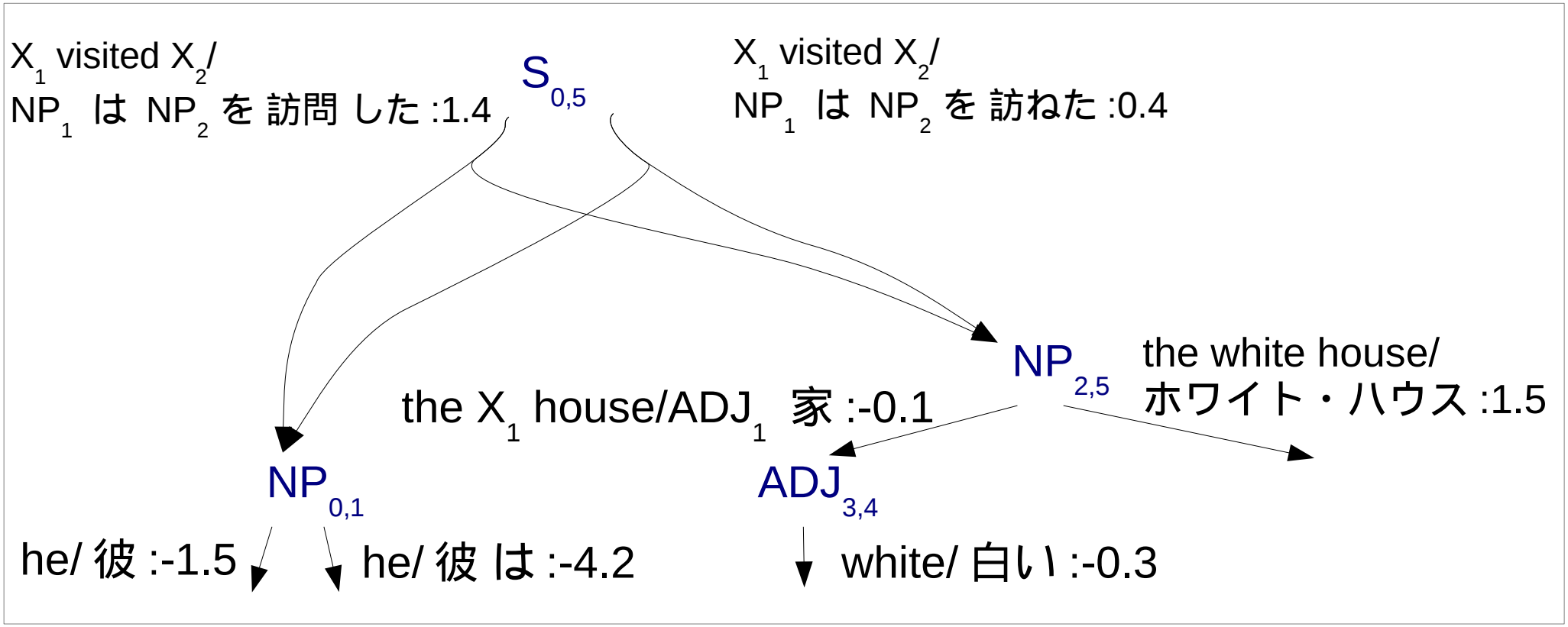
string-to-tree 翻訳

- 目的言語側のみで統語情報を利用
- 階層的フレーズベースとほとんど同じ仕組み
- ルールの目的言語側に句のラベルを付与する

原言語	目的言語	句	スコア
he	彼	NP	-1.5
he	彼 は	NP	-4.2
X_1 visited X_2	NP_1 は NP_2 を 訪問 した	S	1.4
X_1 visited X_2	NP_1 は NP_2 を 訪ねた	S	0.4
the white house	ホワイト ・ ハウス	NP	1.5
the X_1 house	ADJ_1 家	NP	-0.1
white	白い	ADJ	-0.3

string-to-tree 翻訳

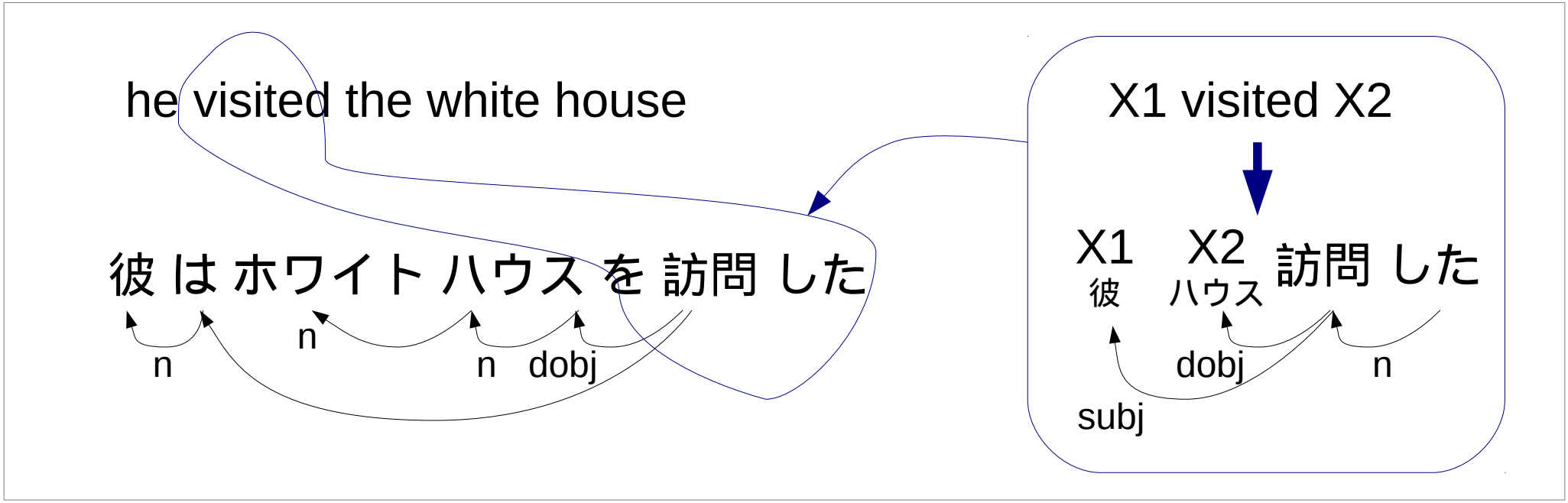
- デコーディングの際は目的言語の句ラベルを考慮



- ルールと合わないラベルのノードを利用しない
  （NP のところに ADJ を入れない）

string-to-dependency 翻訳 [Shen+ 08]

- 出力側の係り受け構造を利用
- 主辞も保持



- 主辞と子供の関係に対する確率を利用

string-to-tree 翻訳の利点と欠点

- + 出力の構文構造がしっかりする
- - 訳出時間が大幅に増加
- - 学習時の構文解析の精度に依存

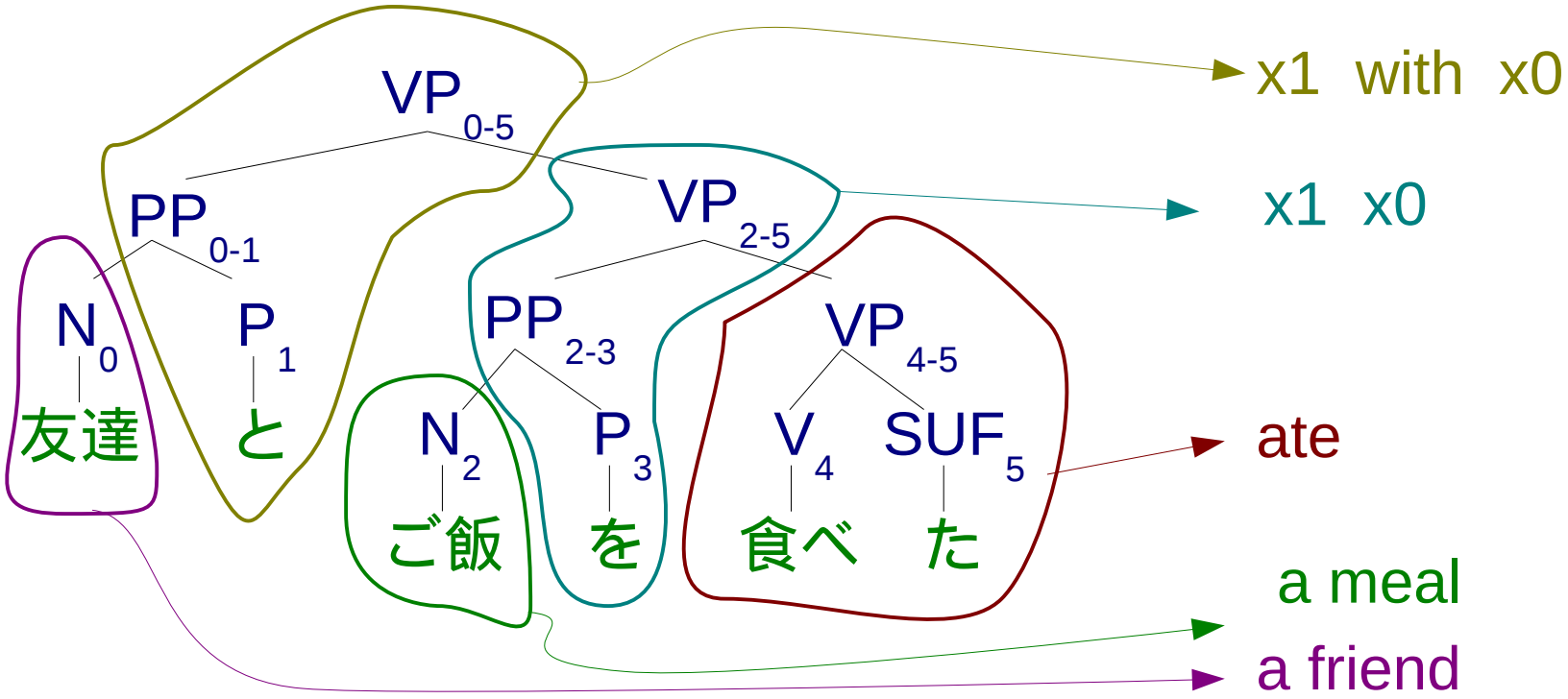
原言語側の構文情報を利用する翻訳

tree-to-string 翻訳

- 原言語側のみに統語情報を利用
- ２種類の方式
  - 同時構文解析＋翻訳：仕組みは string-to-tree と同様
    - ＋構文解析誤りに比較的頑健
    - - 遅い
    - - 並べ替え制限が必要
  - 事前構文解析：事前に解析を行ってから翻訳
    - ＋ 速い
    - ＋ 長距離の並べ替えは問題ない
    - - 解析誤りの影響大

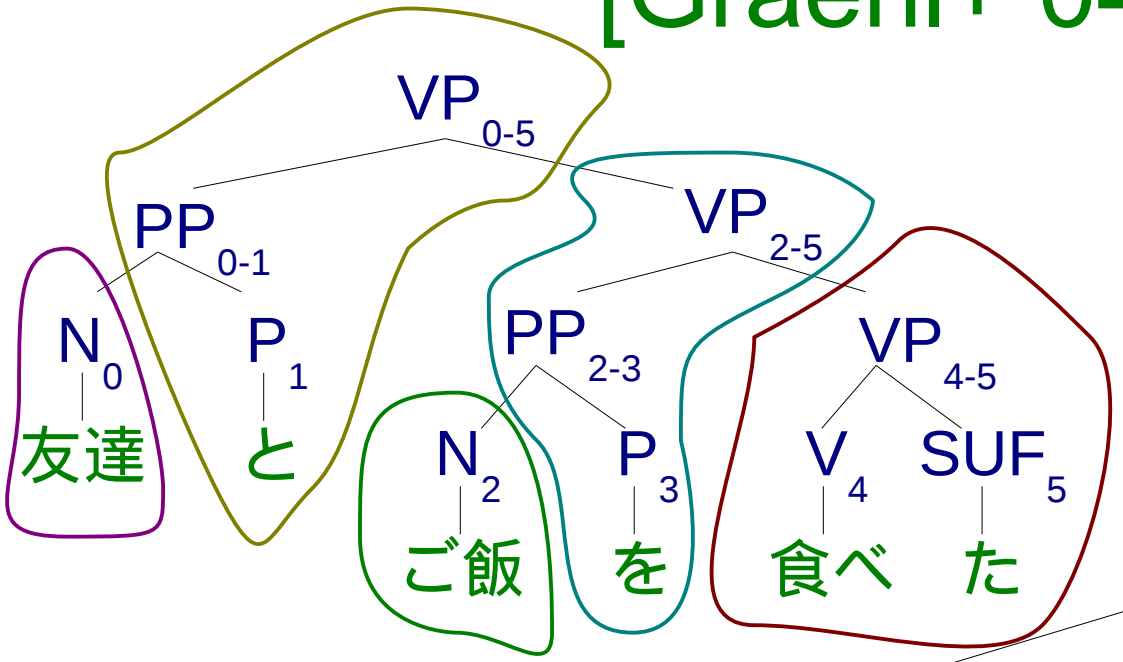
tree-to-string 翻訳 [Graehl+ 04, Liu+06]

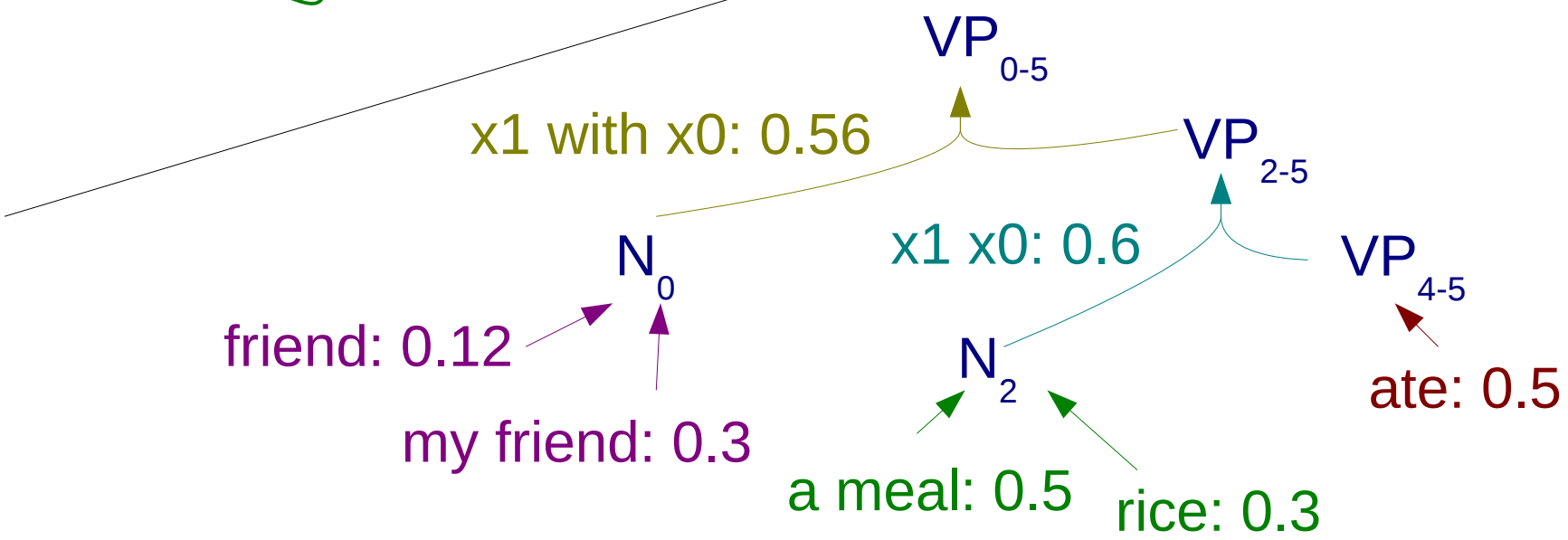
- 構文木上のルールマッチングを行う



tree-to-string 翻訳 [Graehl+ 04, Liu+ 06]

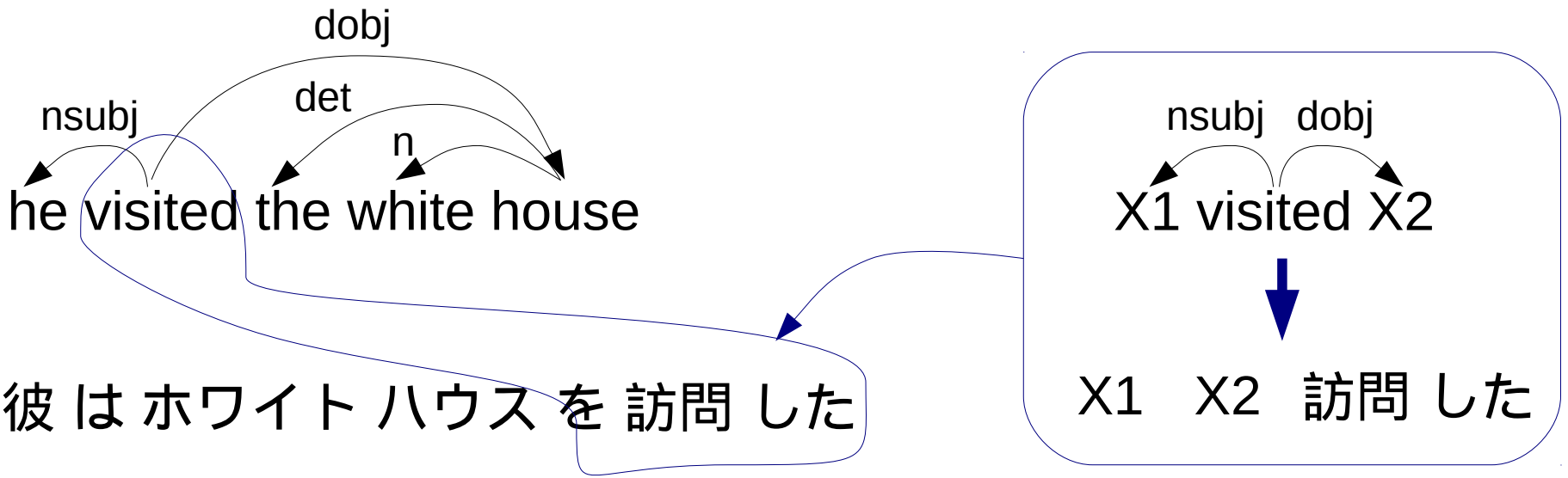
- ルールを表す超グラフを作成
- デコーディングは階層的フレーズベースと類似





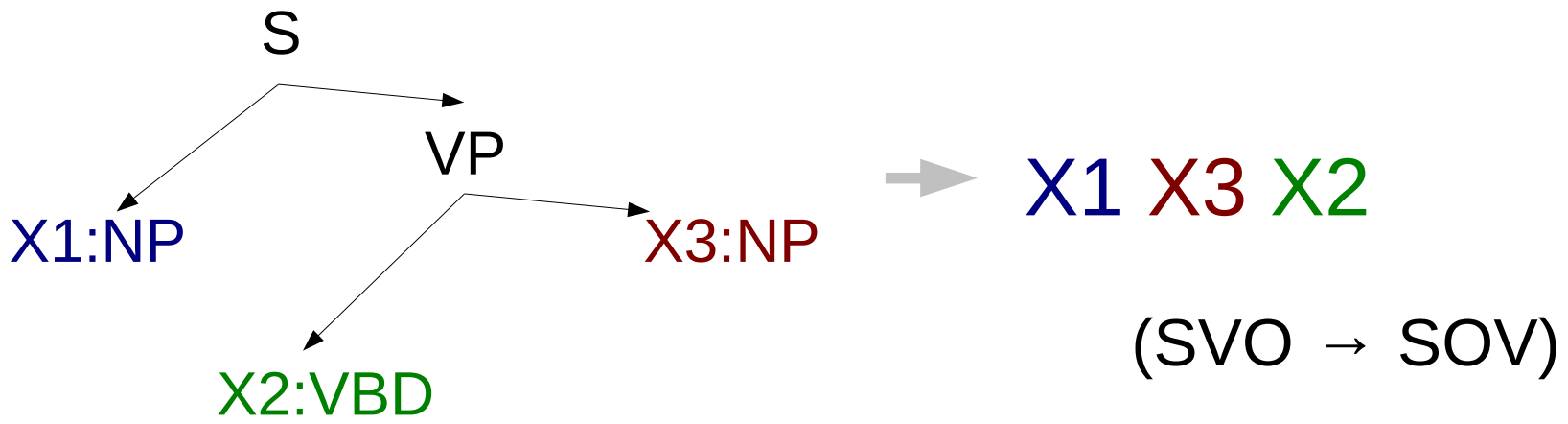
dependency-to-string 翻訳 [Quirk+ 06]

- dependency-to-string 翻訳もある



句構造 vs. 係り受け構造

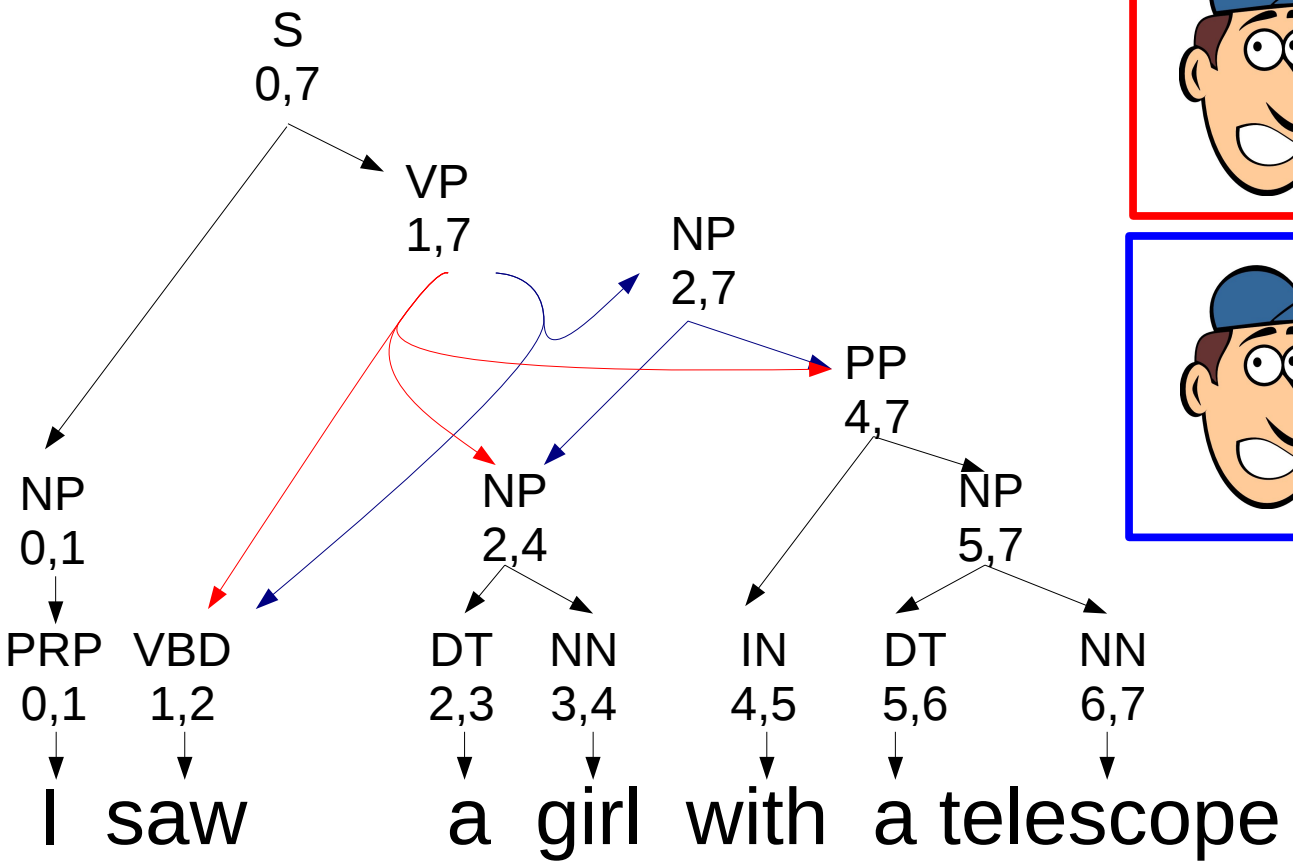
- 句構造：語彙化されていないルールも利用可→一般性

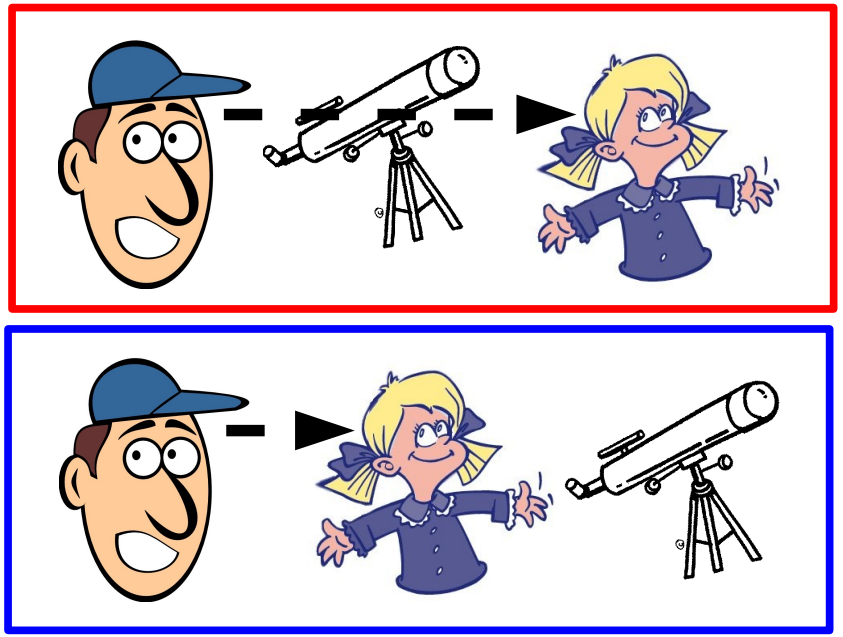


- 係り受け構造：関係のある単語は木上近いところにある→語彙選択に強い？

forest-to-string 翻訳 [Mi+ 08]

- 複数の木を考慮した超グラフを入力

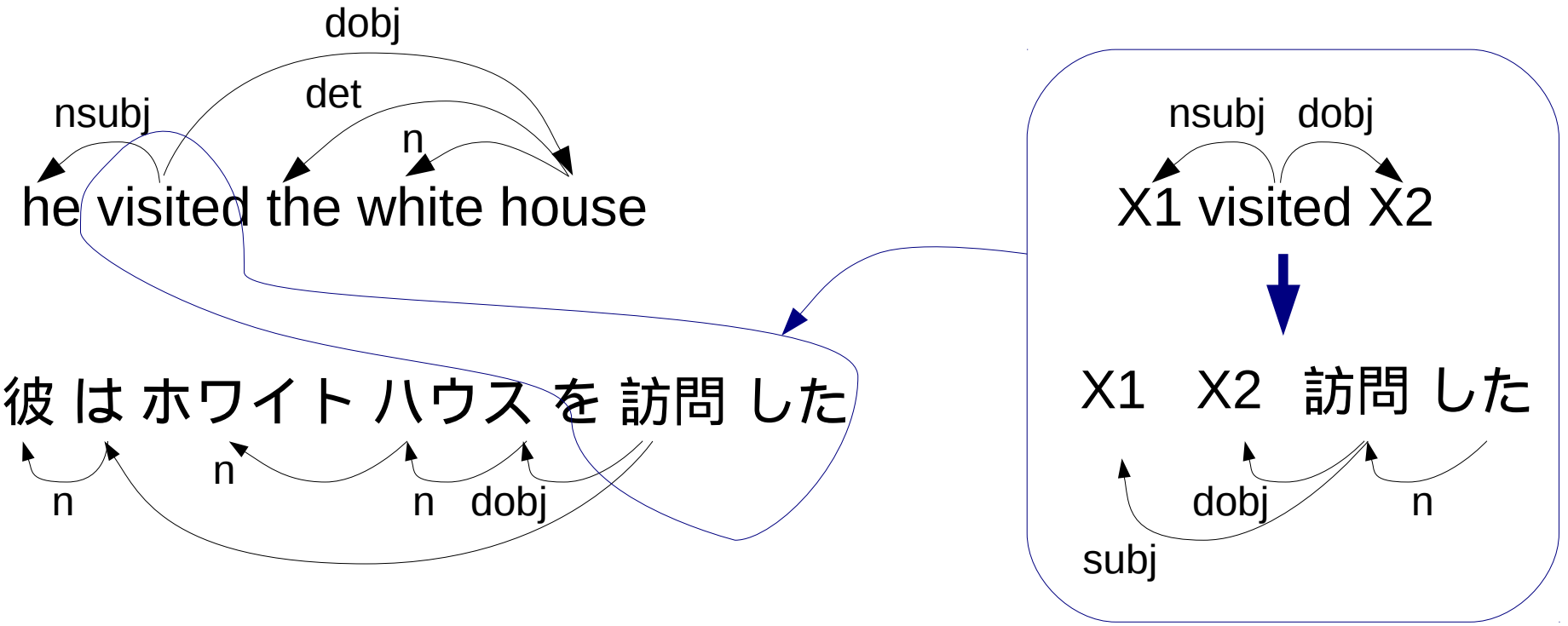




両言語の構文情報を利用する翻訳

dependency-to-dependency 翻訳 [Nakazawa+ 06]

- dependency-to-dependency で両言語に対する係り受けを利用



tree-to-tree 翻訳の利点と欠点

- + 並び替えと目的言語の構造を両方保証
- - 構文解析誤り（学習時・訳出時）に非常に弱い

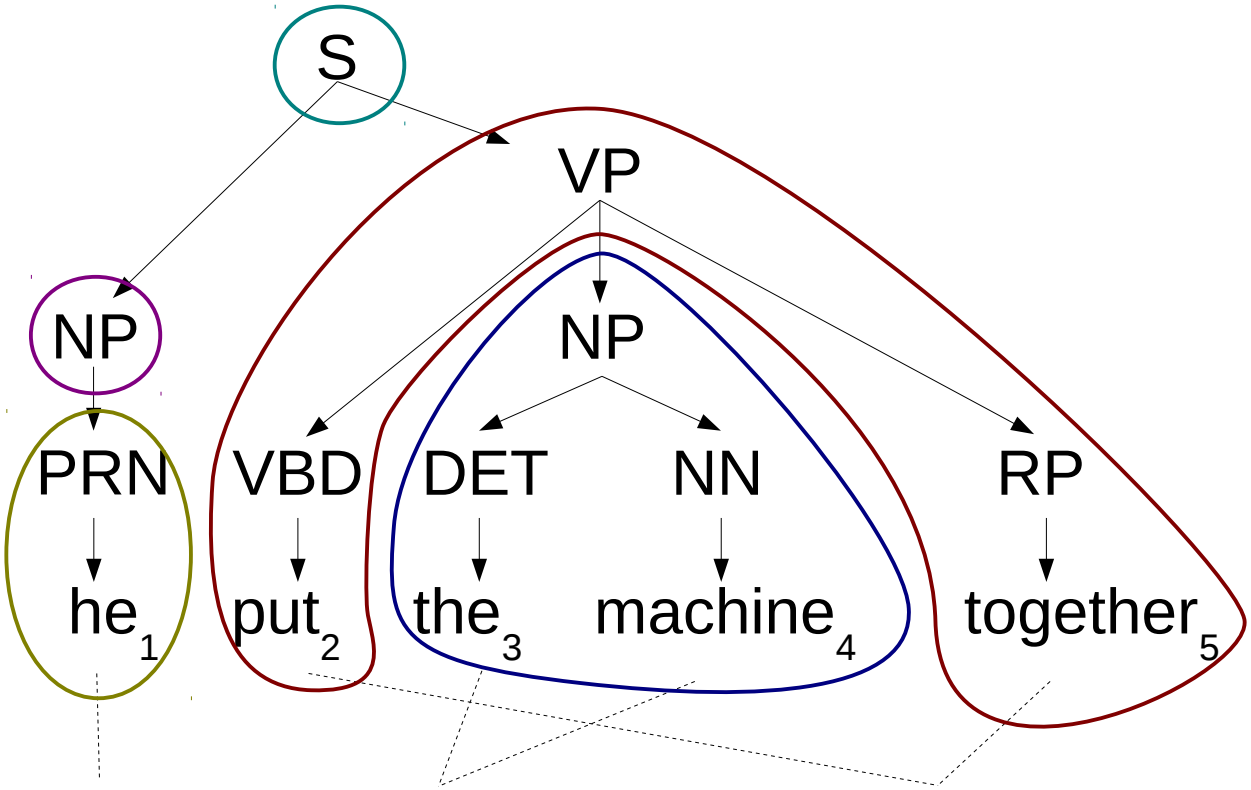
統語情報に基づく翻訳の学習

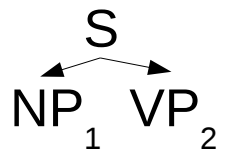
同時文法の学習

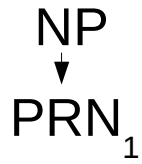
- 構文解析（片方もしくは両方）
- アライメント
- ルール抽出
  - 同時文脈自由文法： Hiero とほぼ同等
  - 同期木置換文法： GHKM アルゴリズム

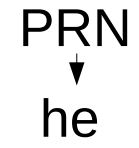
最小ルール [Galley+ 04]

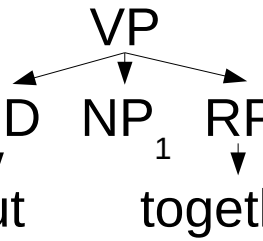
- 「抽出して後の訳出に支障がない最小のルール」

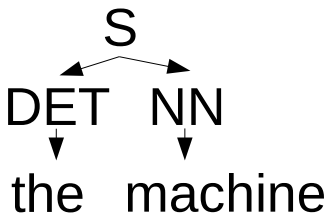


S →  X_1 は X_2

NP →  X_1

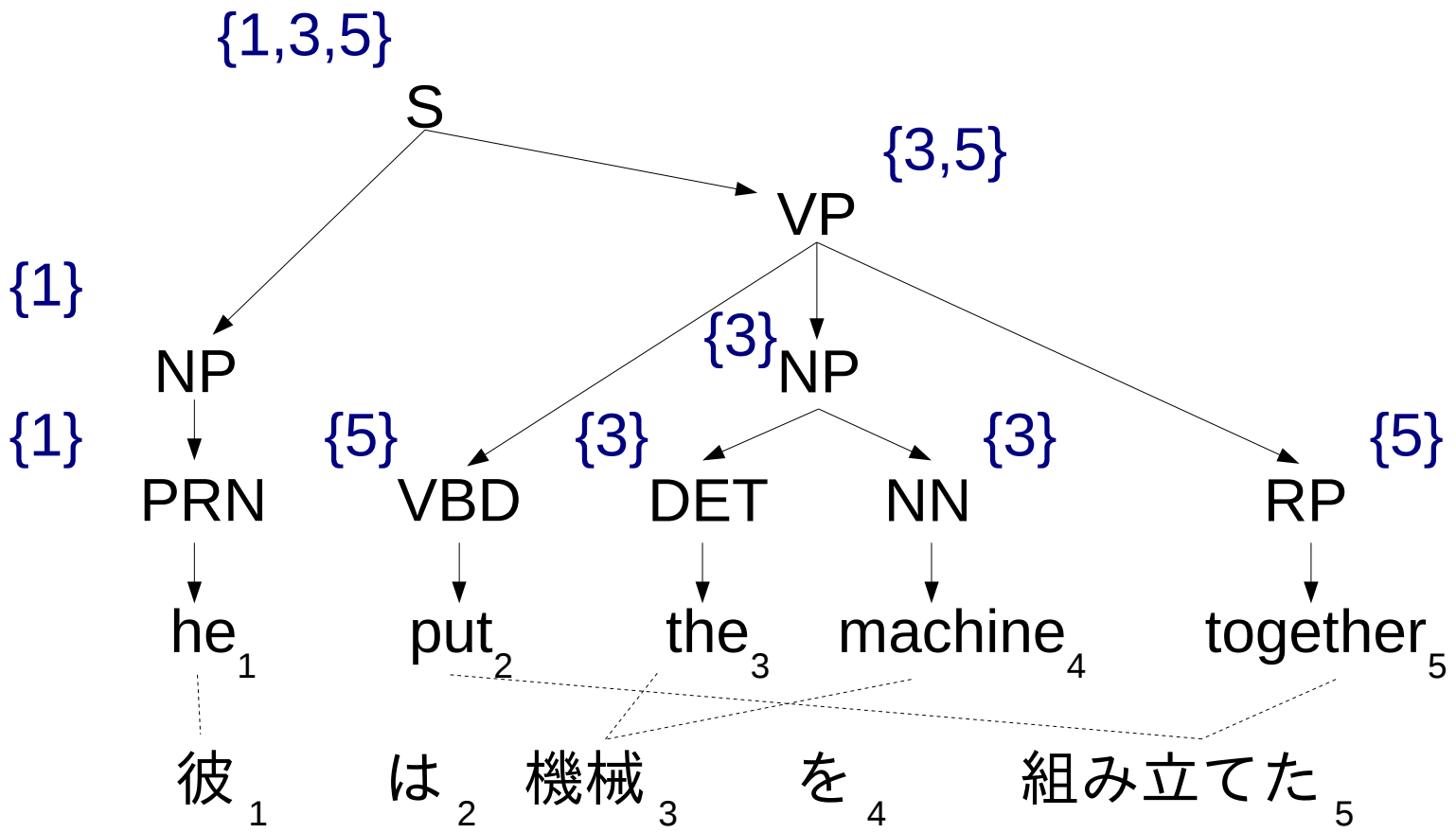
PRN →  彼

VP →  X_1 を 組み立てた

NP →  機械

GHKM アルゴリズム [Galley+ 04]

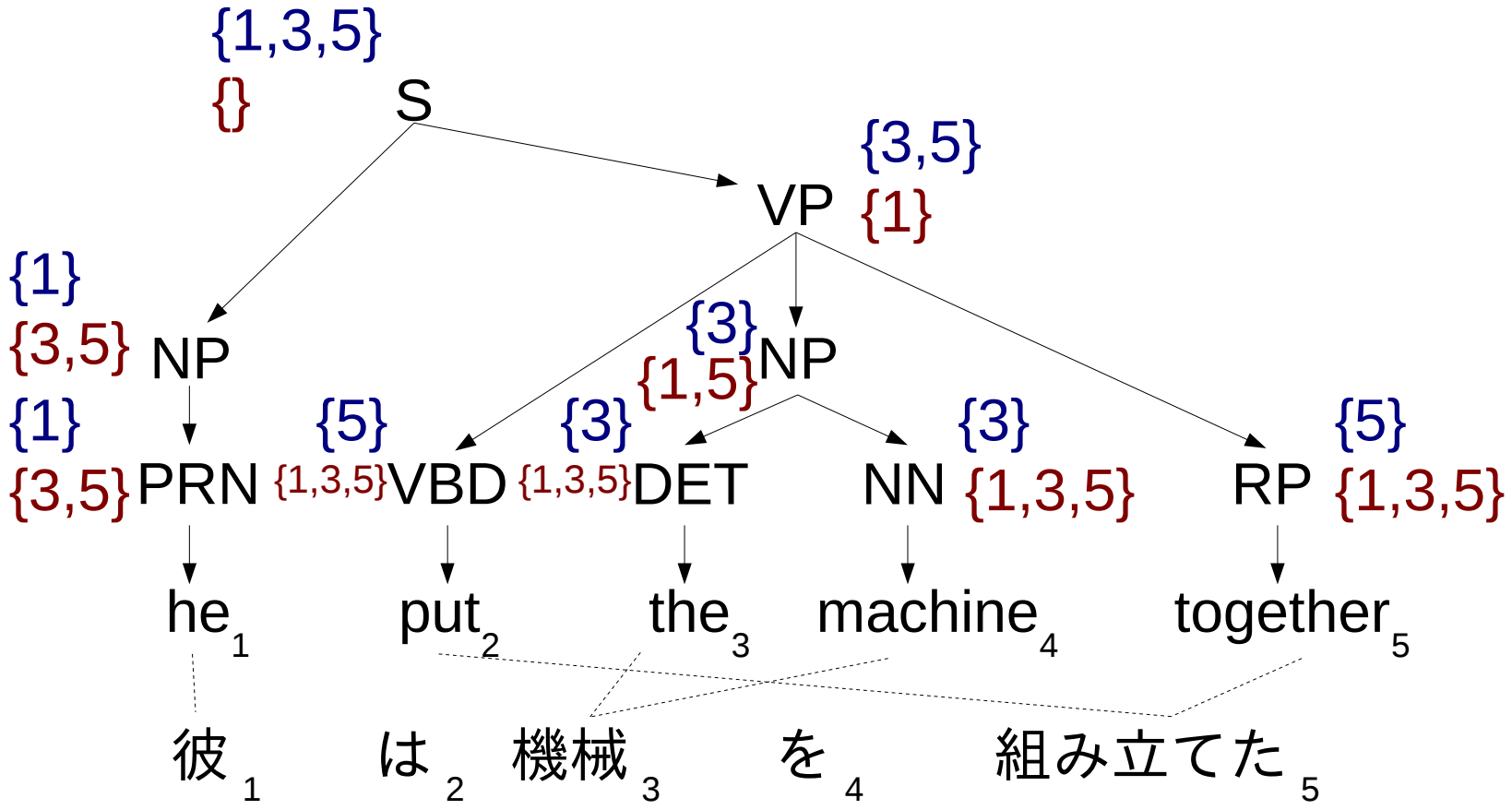
1)ある部分木の対応を表すアライメントスパンを計算



GHKM アルゴリズム [Galley+ 04]

1) ある部分木の対応を表すアライメントスパンを計算

2) ある部分木以外に含まれる補間アライメントスパンを計算

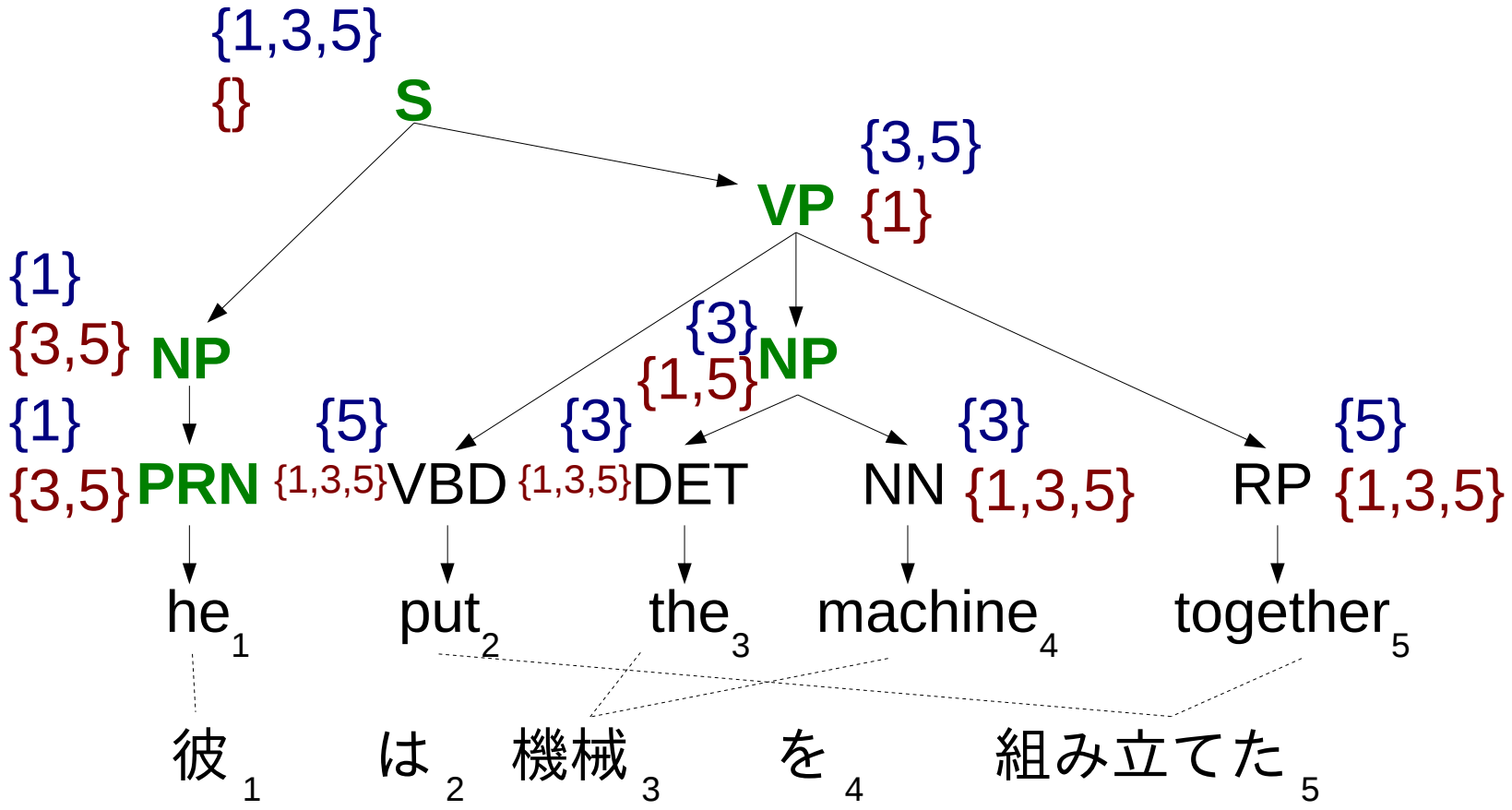


GHKM アルゴリズム [Galley+ 04]

1) ある部分木の対応を表すアライメントスパンを計算

2) ある部分木以外に含まれる補間アライメントスパンを計算

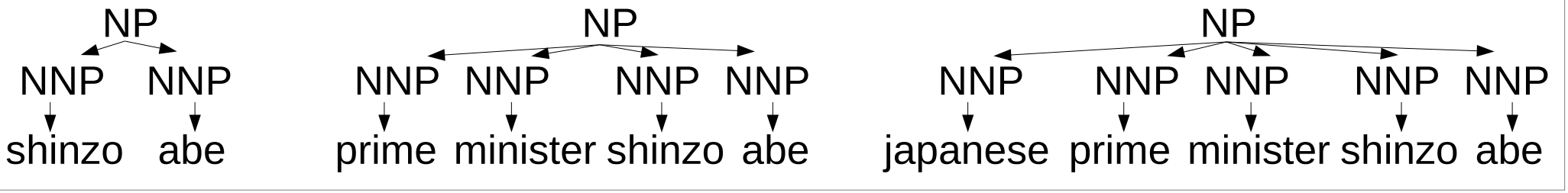
3) スパンと補間スパンが交差しない頂点から始まるルールを抽出



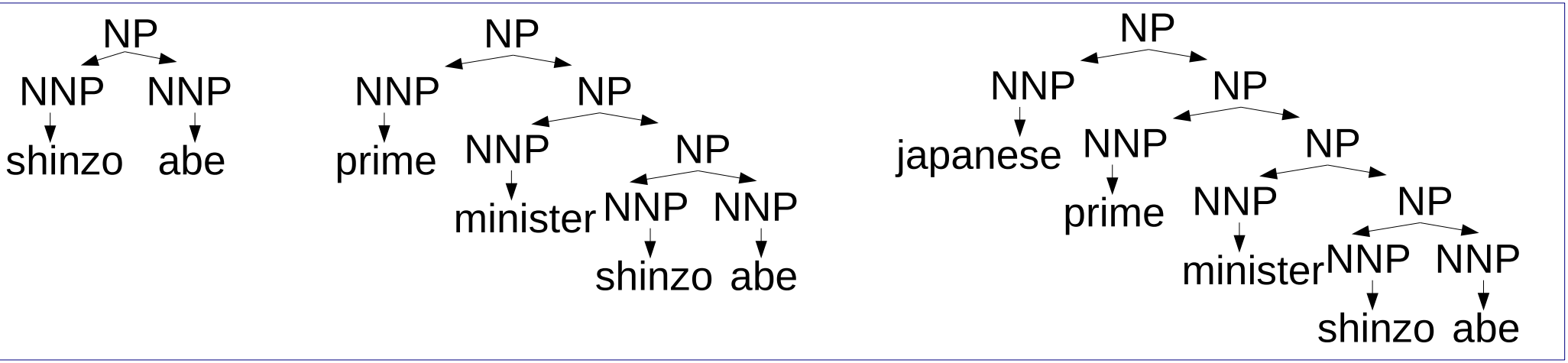
統語情報を用いた翻訳の注意点

２分木化 [Wang+ 07]

- 木の頂点は子供を何個持っても良い



- 通常なら子供の数だけ学習事例が必要…
- 解決策：２分木化



- 右、左、主辞、 CKY など様々

アライメント精度・構文解析精度

- 構文解析精度は重要
- アライメント精度は重要
  - PBMT、 Hiero はそうとは限らない
- 原言語の構文解析森は需要

例：英日特許翻訳における tree-to-string システム

入力	アライメント	構文解析	BLEU	RIBES
木	GIZA++	Stanford	36.23	76.60
木	Nile	Stanford	38.95	78.47
木	Nile	Egret	39.26	79.26
森	Nile	Egret	40.84	80.15

資料・ツール

統語情報を使った翻訳の実装

- 同時文脈自由文法：
  - Moses: 標準的なツールキット
  - cdec: 大規模な学習や最近の最適化を実装
- 木トランスデューサー (森) と同時文脈自由文法：
  - Travatar: 日英の前処理スクリプトなどもある
  - Cicada: 最近の研究を網羅

更に勉強するには


自然言語処理シリーズ 4
機 械 翻 訳
奥村 学 監修
Natural Language Processing Series
コロナ社

6 章

参考文献

- [1] M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, and I. Thayer. Scalable inference and training of context-rich syntactic translation models. In Proc. ACL, pages 961-968, 2006.
- [2] M. Galley, M. Hopkins, K. Knight, and D. Marcu. What's in a translation rule? In Proc. HLT, pages 273-280, 2004.
- [3] J. Graehl and K. Knight. Training tree transducers. In Proc. HLT, pages 105-112, 2004.
- [4] Y. Liu, Q. Liu, and S. Lin. Tree-to-string alignment template for statistical machine translation. In Proc. ACL, 2006.
- [5] H. Mi, L. Huang, and Q. Liu. Forest-based translation. In Proc. ACL, pages 192-199, 2008.
- [6] T. Nakazawa, K. Yu, D. Kawahara, and S. Kurohashi. Example-based machine translation based on deeper NLP. In Proc. IWSLT, pages 64-70, 2006.
- [7] C. Quirk and A. Menezes. Dependency treelet translation: the convergence of statistical and example-based machine-translation? Machine Translation, 20(1):43-65, 2006.