渕一博記念コロキウム
2007年10月20日

制約に基づく言語処理から制約なしの言語処理へ

松本裕治
奈良先端科学技術大学院大学
情報科学研究科

NAIST

簡単な自己紹介

- 1979.4: 電子技術総合研究所入所
  - 推論機構研究室（渕一博室長，翌年より田中穂積室長）
- 1984.9-1985.7: 英国Imperial College滞在研究員（渕さんにKowalski教授への推薦書を書いていただく）
- 1985.9-1987.11
  - ICOT第一研究室（古川康一室長）
- 1988.10: 京都大学
- 1993.4:　奈良先端科学技術大学院大学

NAIST

ICOTでの言語処理研究

- 並列構文解析
  - DCG (Definite Clause Grammars)に基づく構文解析の並列処理
  - PrologおよびGHCによる実装(SAX, PAX)
- 層状ストリームによる並列プログラミング
  - 動的計画法(Dynamic Programming)の並列実装法
  - Prologによる副作用なし・後戻りなしの構文解析(SAX), 形態素解析(LAX)の実装

NAIST

80年代の言語解析

- 単一化文法（制約に基づく文法）
  - 論理文法(Logic Grammars)
    - DCG (Definite Clause Grammars)：文脈自由文法規則（補強項（制約）付き）をPrologの節として直接実行
    - Extraposition Grammars, Gapping Grammars: 関係節などの痕跡(trace)を伴う文法現象等の記述のための拡張
  - HPSG (Head-driven Phrase Structure Grammar)
    - ほとんどの文法情報を語彙に記述(radical lexicalism)
    - 個別の文法規則は存在せず，句の形成を記述する少数のスキーマが存在するだけ

NAIST

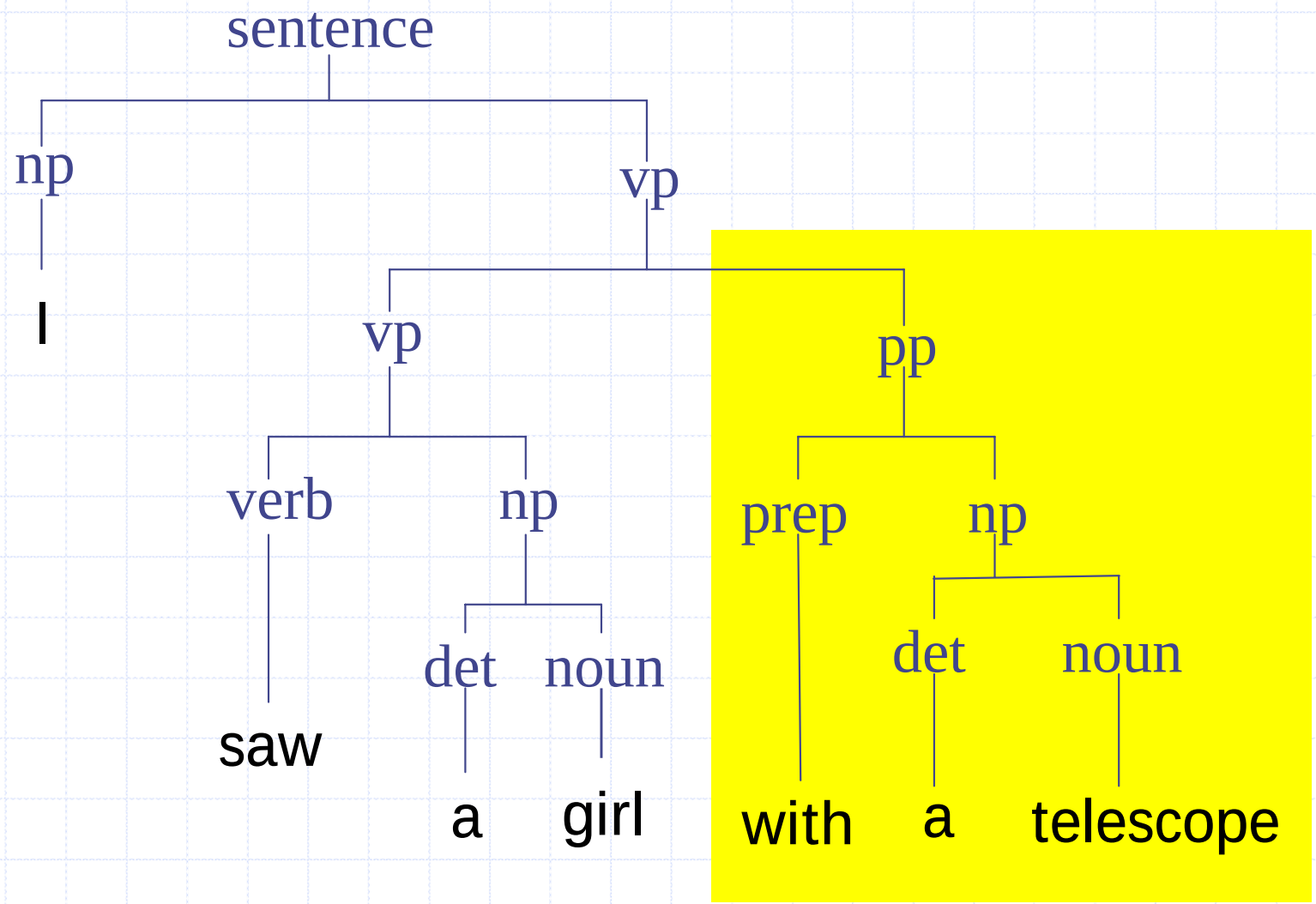
単一化文法(HPSG)の利点・欠点

- 利点
  - 普遍的な演算：2つの句が組み合わさって1つの句を作り上げる際に，単一化演算により一方が他方の関数として機能する
  - 文法の記述(語彙情報＋原則)と処理の独立性
- 欠点
  - 曖昧性爆発：1つの文に対して，文法的に正しい解析結果が多数得られる
  - 脆弱性：文法誤りを持つ文(あるいは，想定外の文法現象)に対して，処理が破綻する

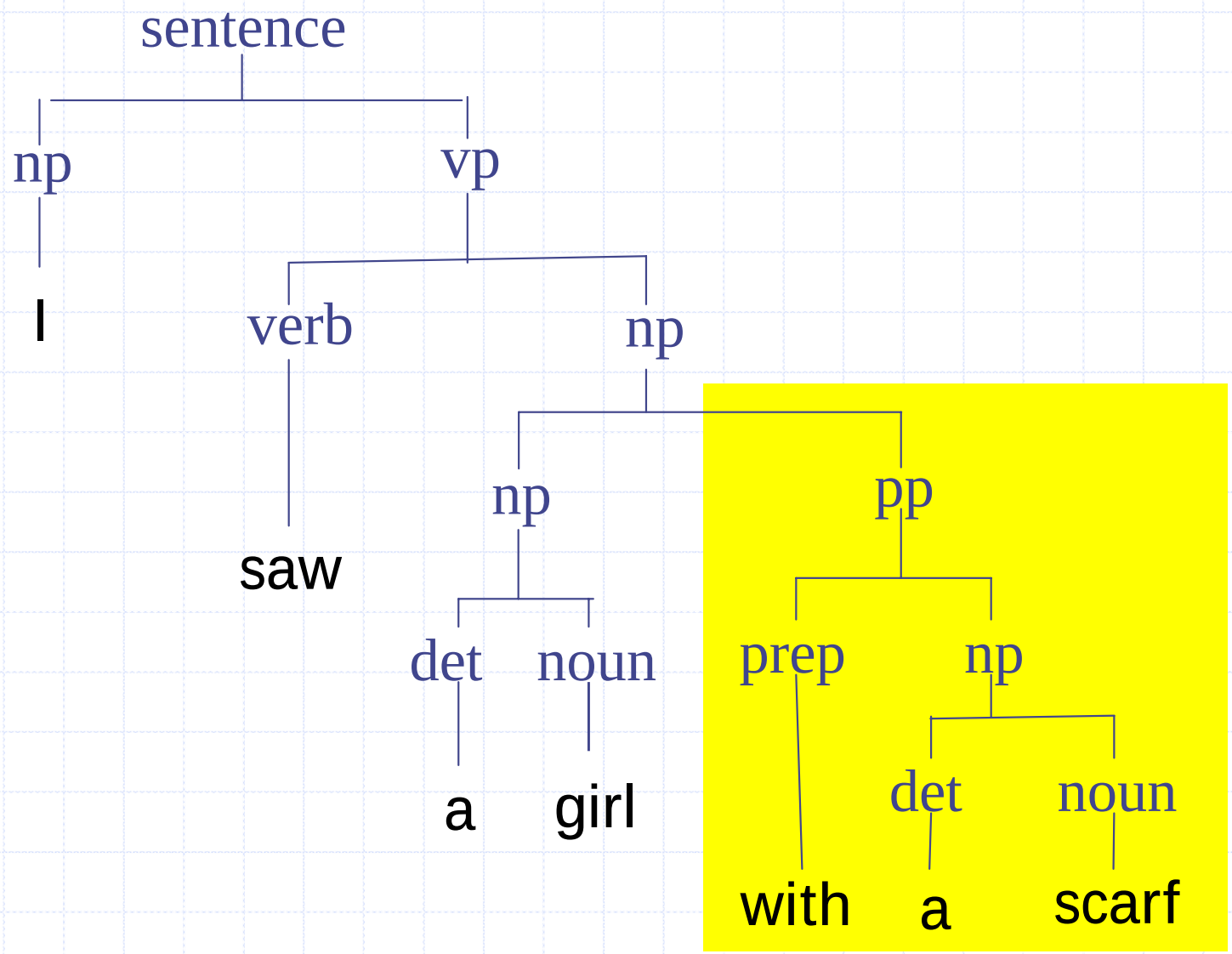
曖昧性の問題

頑健性の問題

統語的曖昧性の例


sentence
np
vp
I
vp
pp
verb
np
prep
np
det
noun
det
noun
saw
a
girl
with
a
telescope

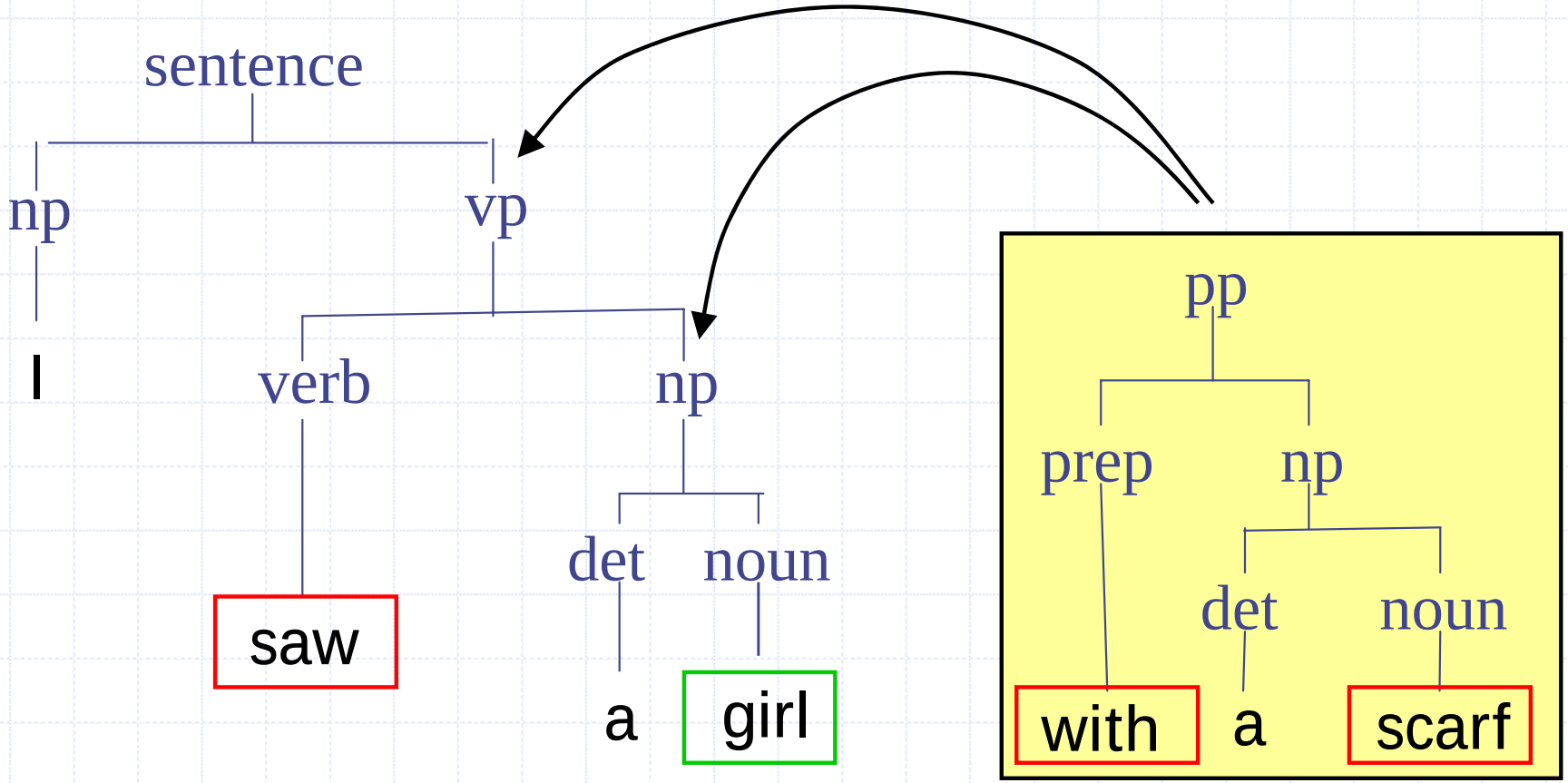
統語的曖昧性の例


sentence
np
vp
I
verb
np
saw
np
pp
det
noun
prep
np
a
girl
det
noun
with
a
scarf

90年代前半の言語解析

- コーパス(大規模言語データ)の出現
  - 統計的機械翻訳(Statistical Machine Translation)– IBMグループ：英仏翻訳
  - 隠れマルコフモデルによる品詞タグ付けプログラム(形態素解析)
  - 確率文脈自由文法(Probabilistic Context-free Grammar)
    - 各文法規則(VP→V NP, VP→V NP PP など)に確率値を与え，文を構成する文法規則の積が最大になる構文木を求める
    - Inside-Outsideアルゴリズム(Expectation Maxmization法の一種)により，自動推定可能

統語的曖昧性の例


sentence
np
vp
I
verb
np
saw
det
noun
a
girl
pp
prep
np
det
noun
with
a
scarf

NAIST

初期の統計モデルの問題点

- 品詞や句を確率事象の単位とした条件付確率を用いる：単語の情報を考慮しにくい
- 同じ品詞列や句の組み合わせを持つ文でも異なる形で解析すべき例（telescopeの例）がある
  - 次の例は同じ品詞列（名詞，助詞，動詞，名詞，助詞，動詞）を持つが，異なる構造を持つ
    - 双眼鏡で泳ぐ子供を監視した　（双眼鏡で→監視した）
    - 海で泳ぐ子供を監視した　（海で→泳ぐ）
- 細かい粒度の情報（単語や単語の接頭・末尾の文字列など）を取り込むことのできる統計モデルへ：
  - Exponential models (最大エントロピー法，Conditional Random Fieldsなど）
  - Maximum margin methods (Support Vector Machines, Boosting など)

NAIST

統計学習と言語処理タスクの分類

学習タスクの分類	言語処理タスクの例
予測	言語モデル，語の類似度
分類	文書分類，語義曖昧性解消(WSD)，用語の意味分類，照応解析（代名詞の指示物の同定）
系列タグ付け	分かち書き，品詞タグ付け，基本句チャンキング，固有表現抽出，統語解析(句構造解析，係り受け解析)
変換	統計的機械翻訳，機械翻訳規則獲得， 対訳文アラインメント，
マイニング	コロケーション，クラスタリング，語彙知識獲得，素性選択

NAIST

語彙情報を利用した統計的統語解析

- 句構造解析
  - 決定木の利用(Magerman 95)
  - 句の主辞の共起確率を用いた句構造解析(Collins 96)
  - 最大エントロピー法の利用 (Ratnaparkhi 97, Charniak 00)
- 係り受け解析
  - SVMの利用したchunking cascade (Kudo & Matsumoto 02, Yamada & Matsumoto 03)
  - Memory-based Reasoningを利用したShift-Reduce法 (Nivre 03)
  - Maximum Spanning Tree algorithmとマージン最大化の利用 (McDonald 05)
- その他：re-ranking手法など

Japanese Dependency Parsing by CaboCha

NAIST

自民党は20日、衆院選小選挙区の第5次公認候補4人を発表した。

input sentence

解析結果

```
        <ORGANIZATION>自民党</ORGANIZATION>は---------D
                                <DATE>２０日</DATE>、-------D
    <ORGANIZATION>衆院</ORGANIZATION>選小選挙区の-D    |
                                        第５次公認候補-D |
                                                ４人を-D
                                              発表した。
EOS
* 0 5D 0/1 4.13561692
自民党      ジミントウ    自民党      名詞-固有名詞-組織                    B-ORGANIZATION
は          ハ           は          助詞-係助詞                  0
* 1 5D 2/2 2.98545849
２          ニ           ２          名詞-数                      B-DATE
０          ゼロ         ０          名詞-数                        I-DATE
日          ニチ         日          名詞-接尾-助数詞                       I-DATE
、          、           、          記号-読点                  0
* 2 3D 4/5 0.96527839
衆院        シュウイン   衆院        名詞-固有名詞-組織                          B-ORGANIZATION
選          セン         選          名詞-接尾-一般                    0
小          ショウ       小          接頭詞-名詞接続                      0
選挙        センキョ     選挙        名詞-サ変接続                  0
区          ク           区          名詞-接尾-地域                     0
の          ノ           の          助詞-連体化                0
* 3 4D 4/4 1.49313142
第          ダイ         第          接頭詞-数接続                  0
５          ゴ           ５          名詞-数                  0
次          ジ           次          名詞-接尾-助数詞                   0
公認        コウニン     公認        名詞-サ変接続                  0
候補        コウホ       候補        名詞-一般                    0
```

dependency tree

文節まとめ上げ

固有表現のタグ付け

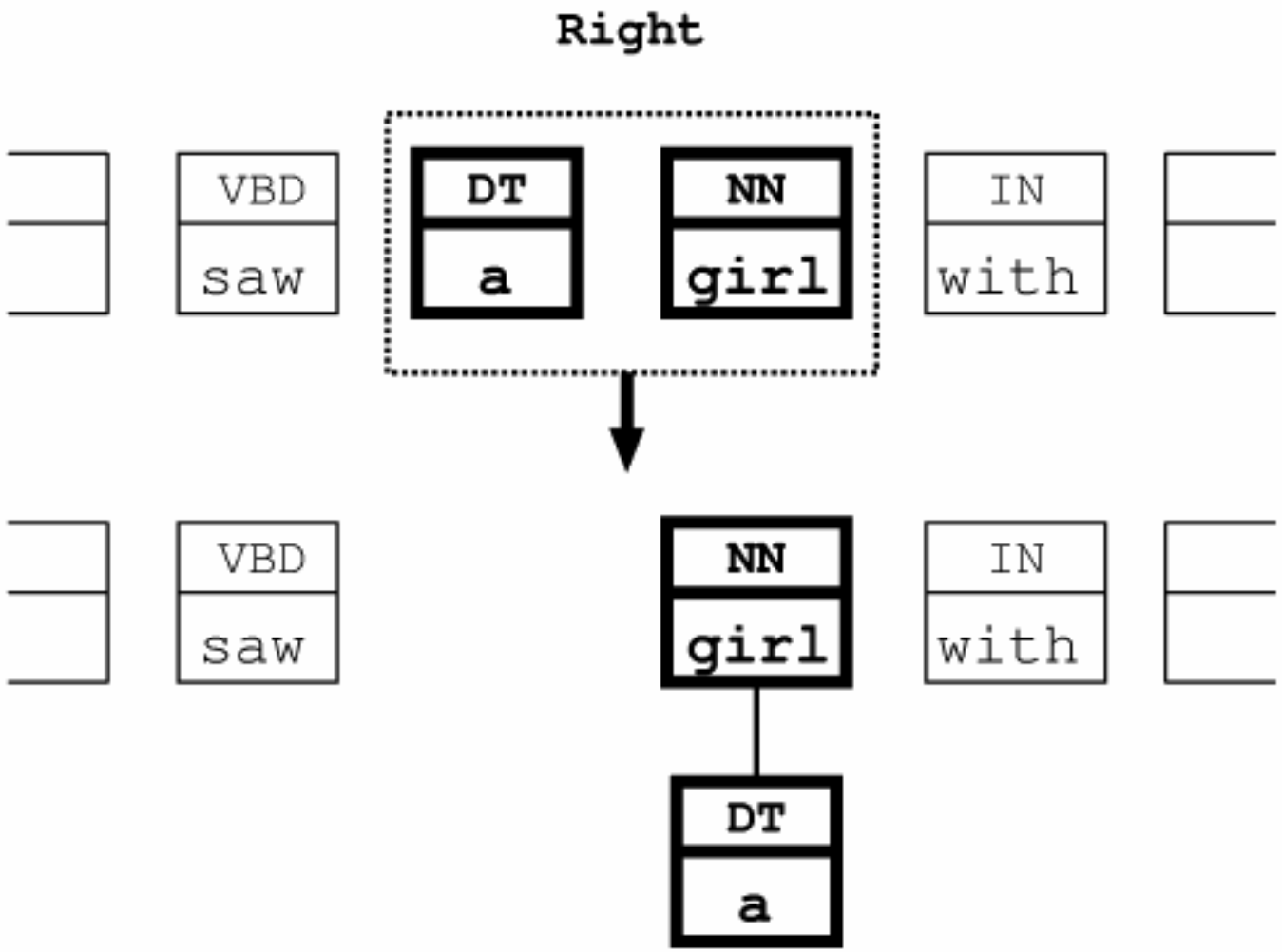
英語に対する決定性の単語係り受け解析

[Yamada & Matsumoto 03]

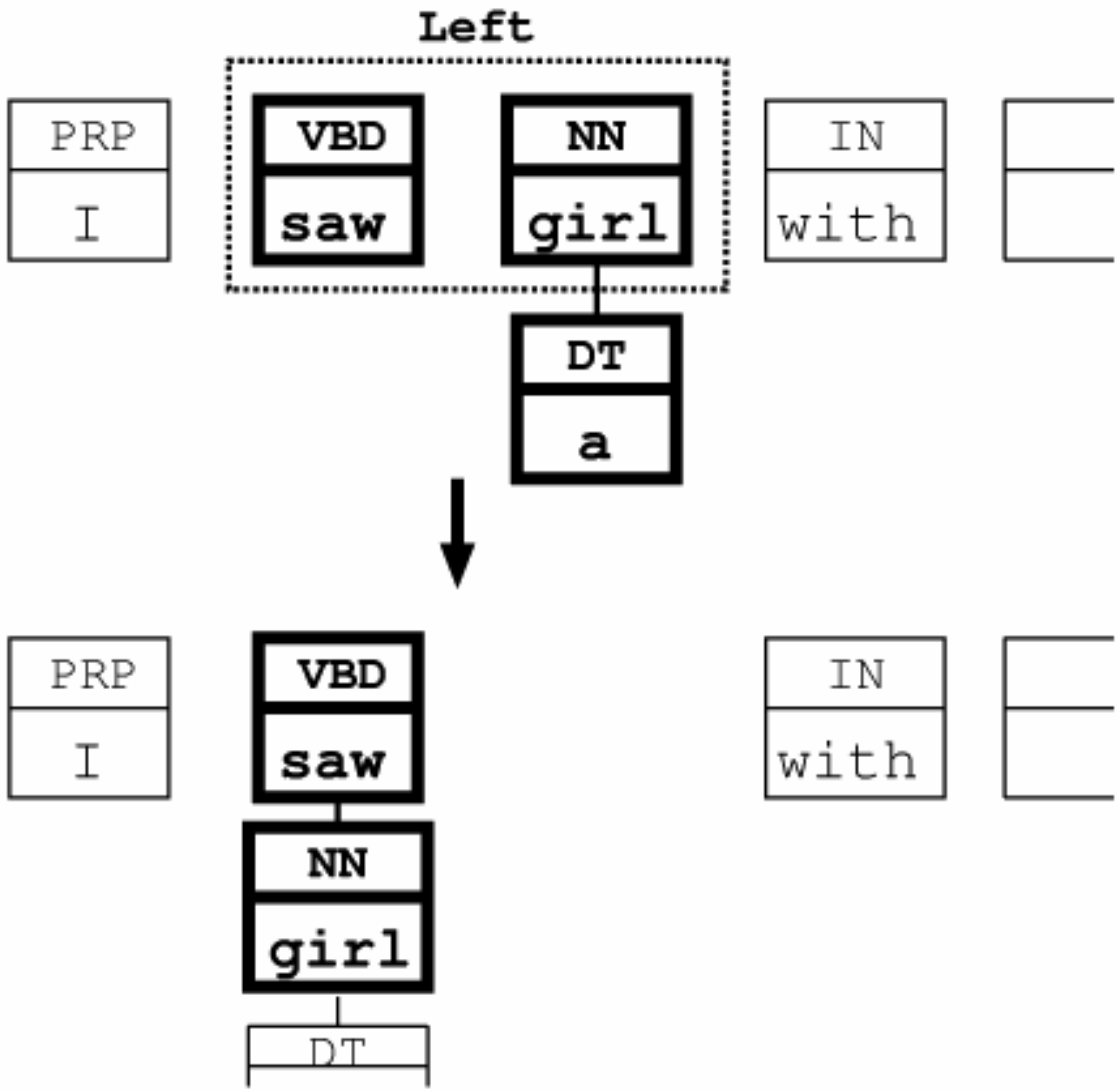
- 各状況で3つのアクションを考える:
  - Right: 2つの連接する単語の間に，左から右への係り受け関係がある．左の語を右の語へ掛けて，消去する
  - Left: 2つの連接する単語の間に，右から左への係り受け関係がある．右の語を左の語へ掛けて，消去する
  - Shift: 現在の2つの単語の間には，係り受け関係を決めずに，対象を一つ右へずらす
    - この状況では，実は2つの可能性がある:
    - 1．2つの単語の間に依存関係がない
    - 2．2つの単語の間に本当は依存関係があるかも知れないが，この場で決めることを避けて，処理対象を右へ移す．（次のラウンドで決定する）

      この場合を，上の場合と区別して “Wait” というアクションをたてることもできる
- この処理を文頭から順に右へ向かって適用し，文末まで来れば，文頭へ向かって処理を繰り返す．文全体が一つの依存構造木になれば終了

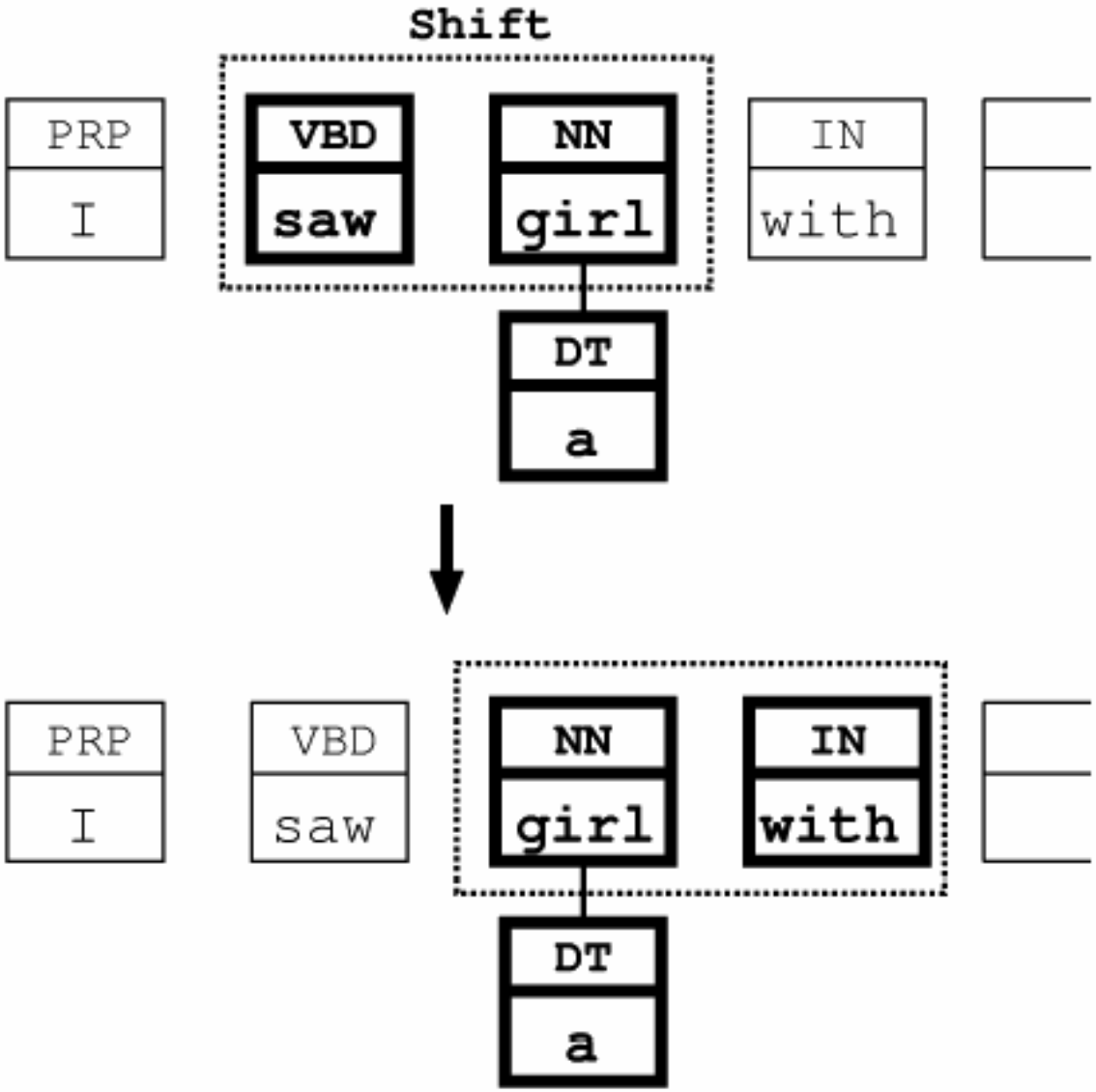
Right action



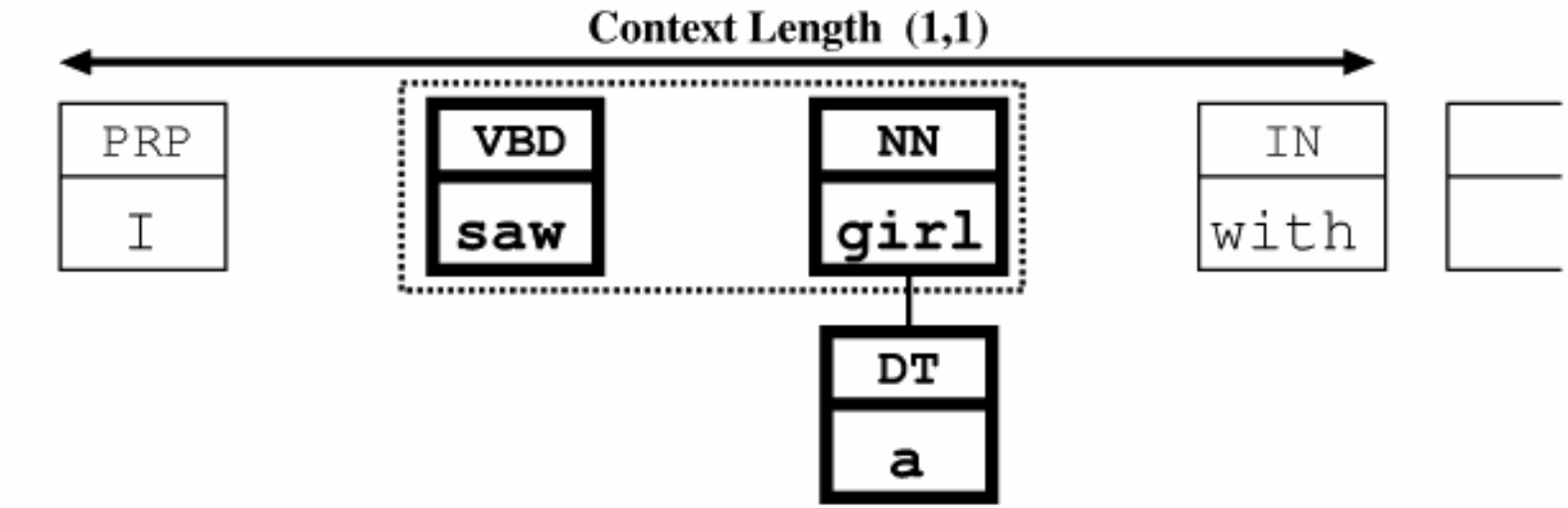
Left action


Left
PRP
I
VBD
saw
NN
girl
IN
with
DT
a
PRP
I
VBD
saw
NN
girl
DT
IN
with

Shift action



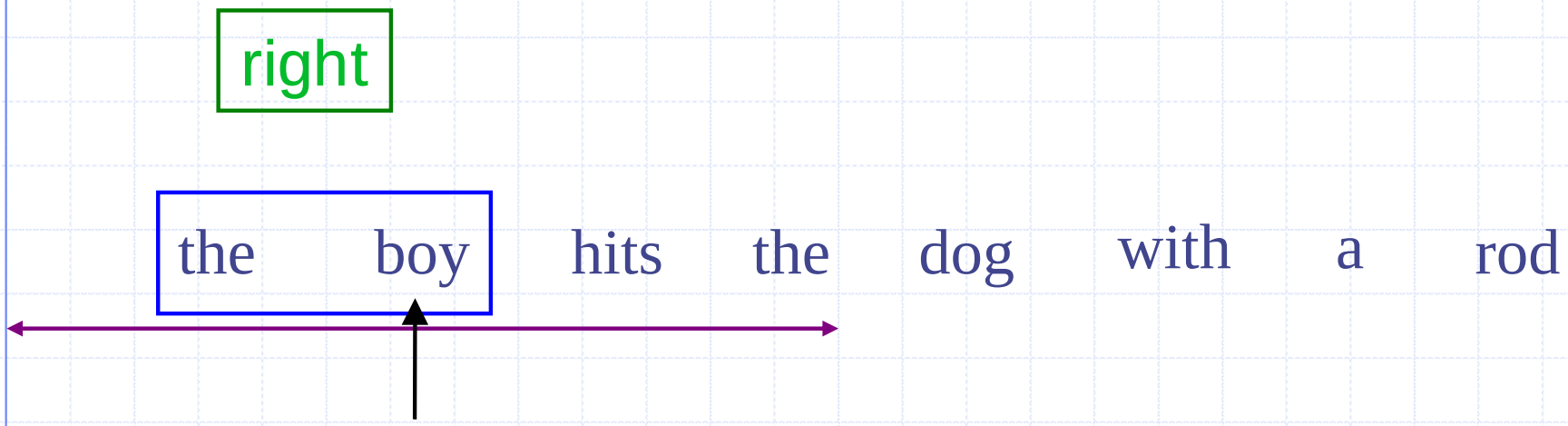
学習に用いられる素性(属性)



-1.lex.I
-1.pos.PRP
0-.lex.saw
0-.pos.VBD
0+.lex.girl
0+.pos.NN
0+.pos(L).DT
0+.lex(L).a
+1.lex.with
+1.pos.IN

3 class問題 (right, left, shift) または
4 class問題 (right, left, shift, wait)を解くために
SVM(Support Vector Machines)を用いる

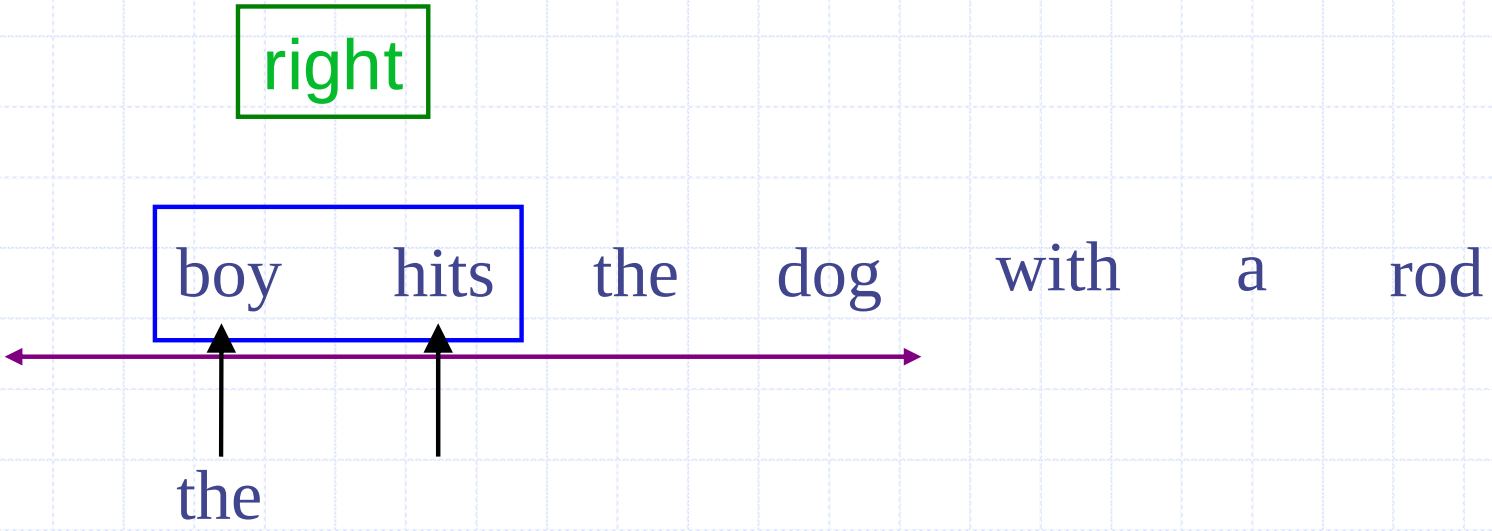
Yamada法による英語の係り受け解析の例



処理すべき単語対

参照文脈

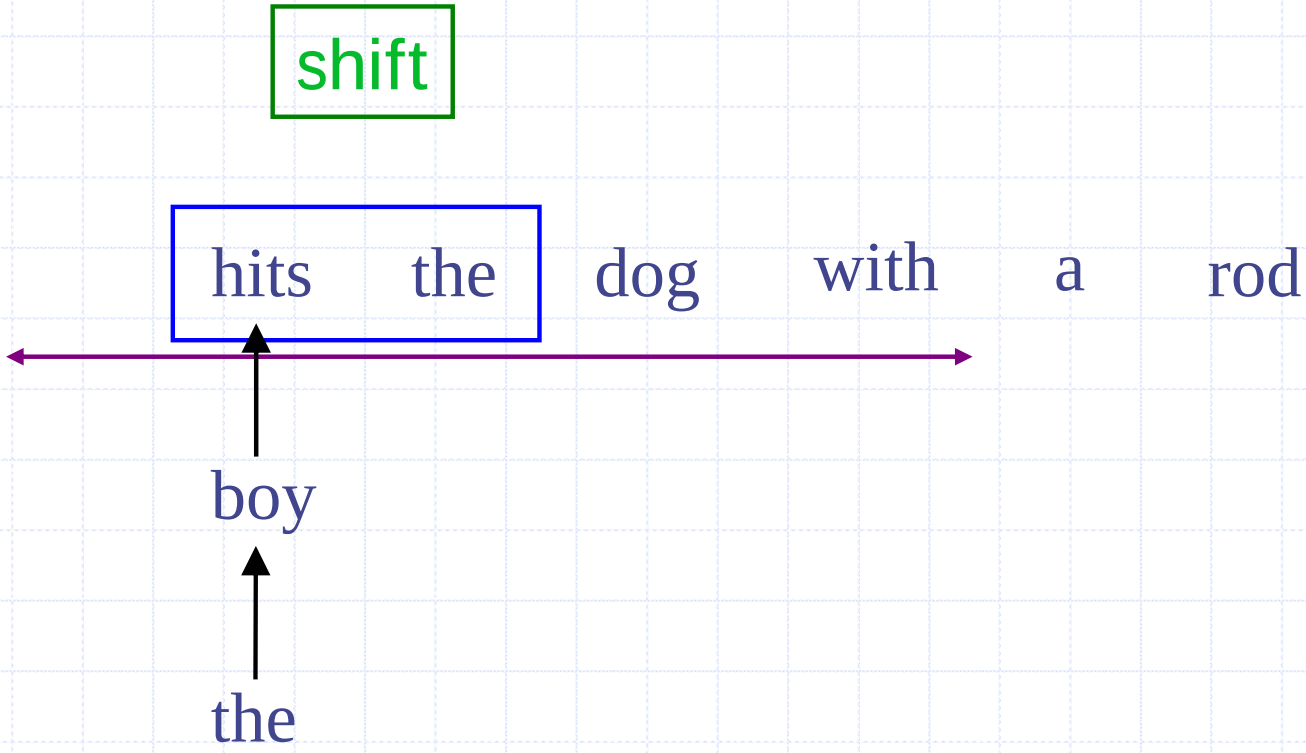
Yamada法による英語の係り受け解析の例



処理すべき単語対

参照文脈

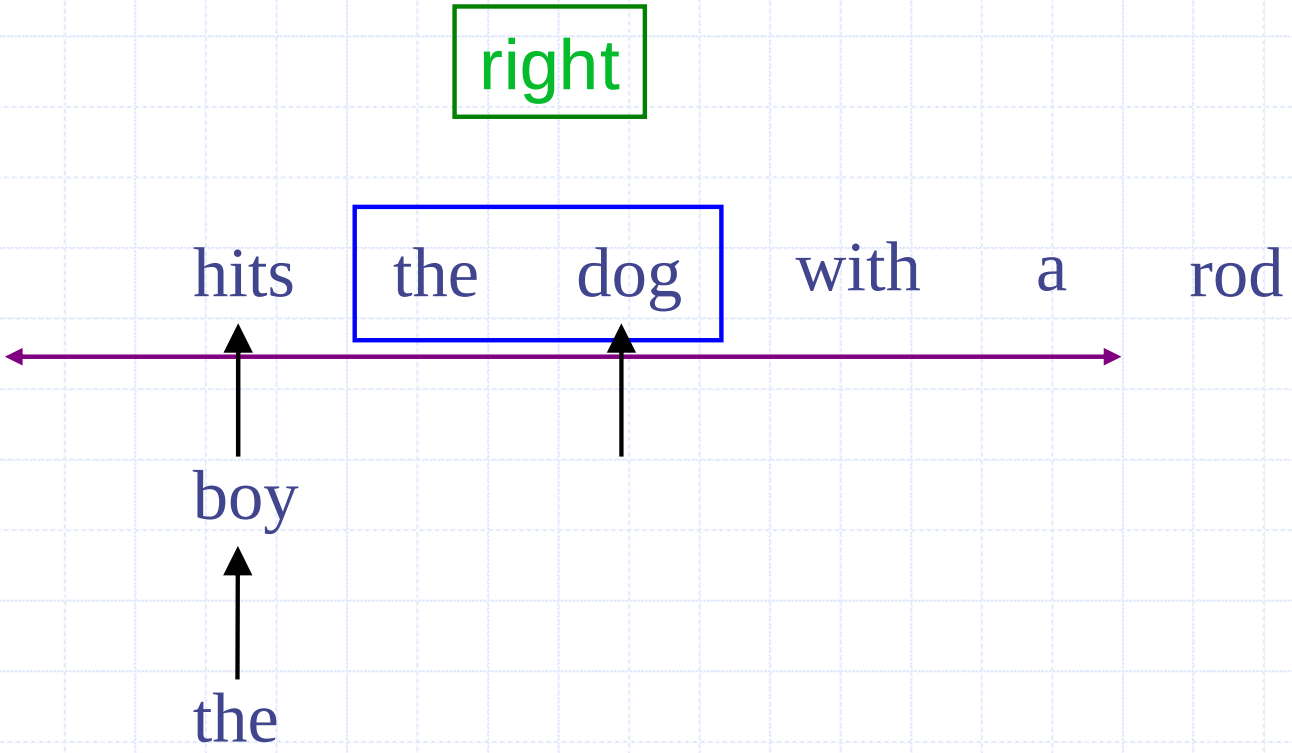
Yamada法による英語の係り受け解析の例



処理すべき単語対

参照文脈

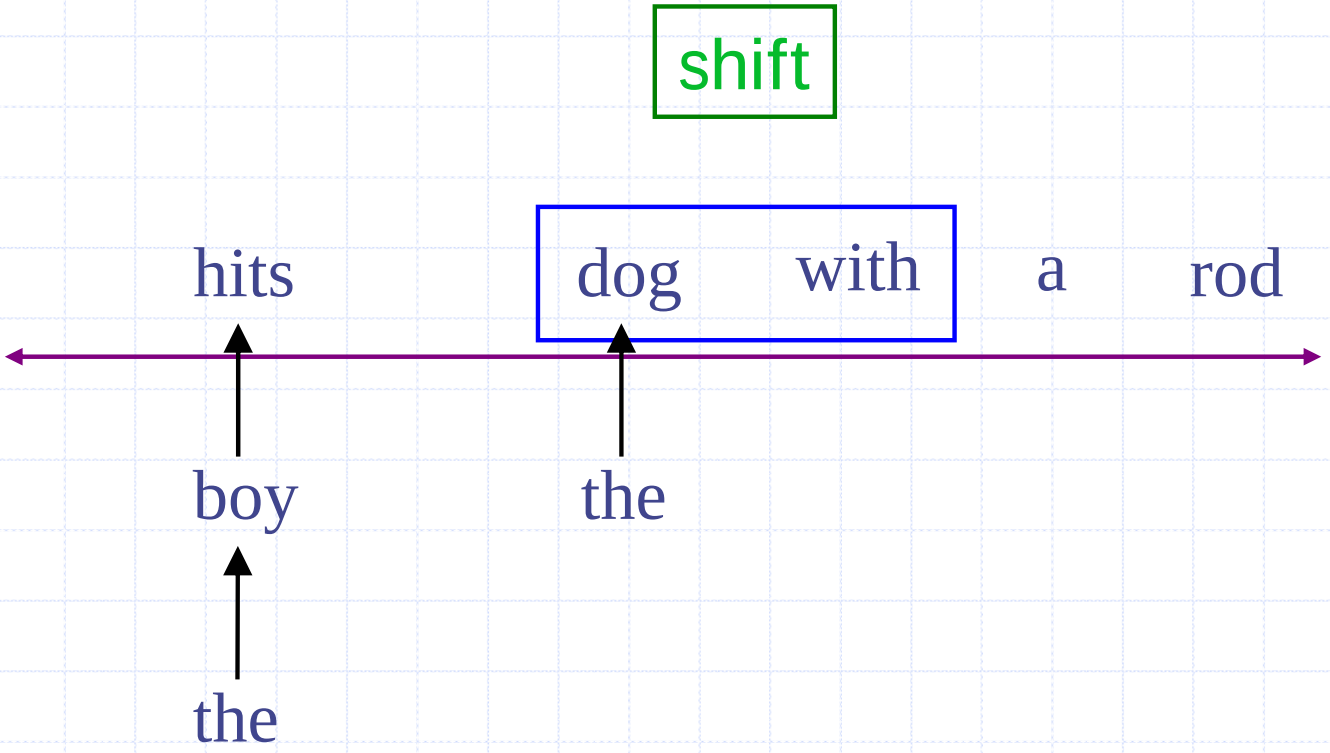
Yamada法による英語の係り受け解析の例



処理すべき単語対

参照文脈

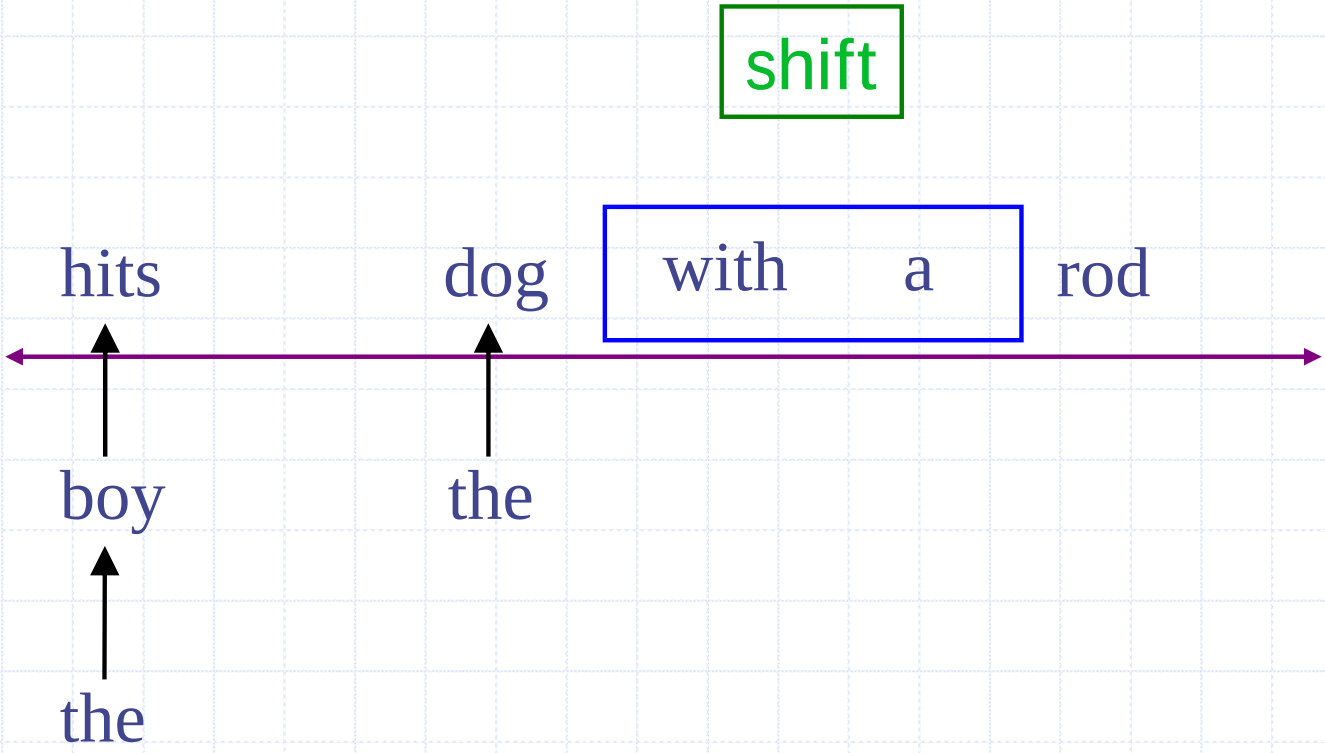
Yamada法による英語の係り受け解析の例



処理すべき単語対

参照文脈

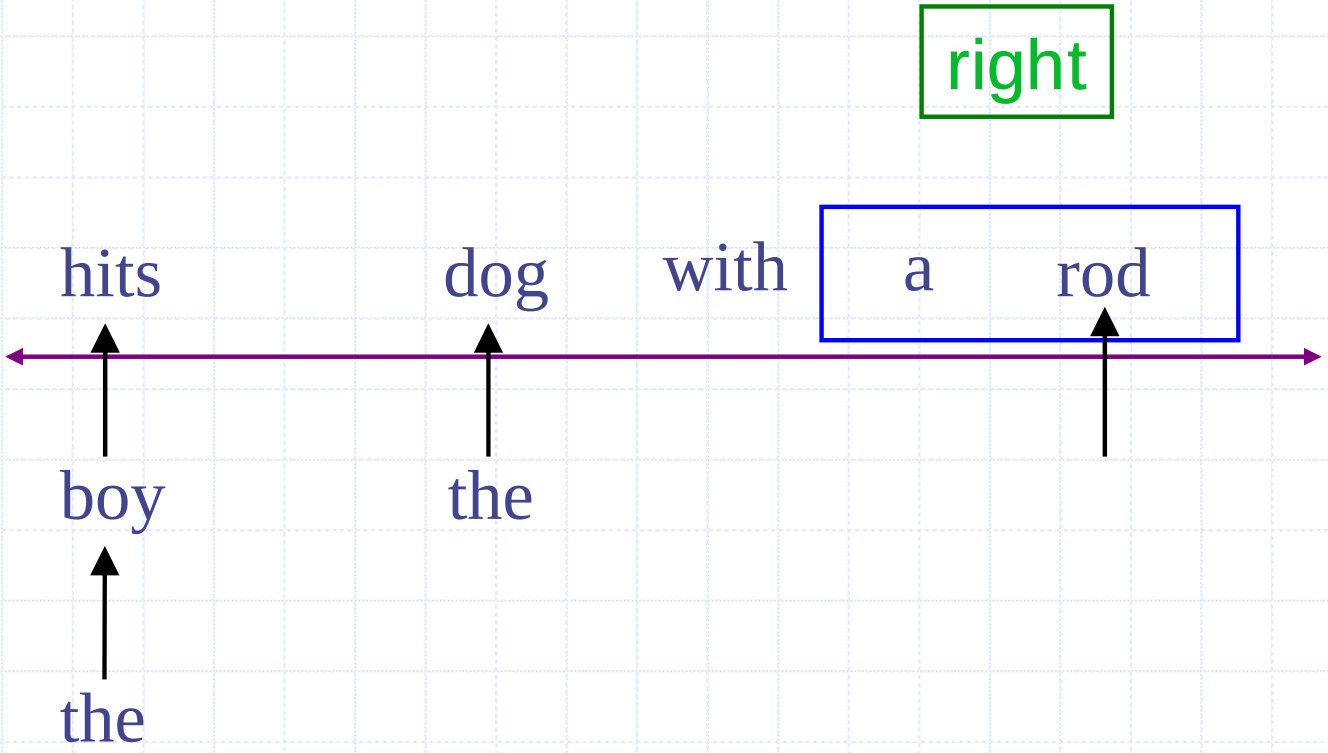
Yamada法による英語の係り受け解析の例



処理すべき単語対

参照文脈

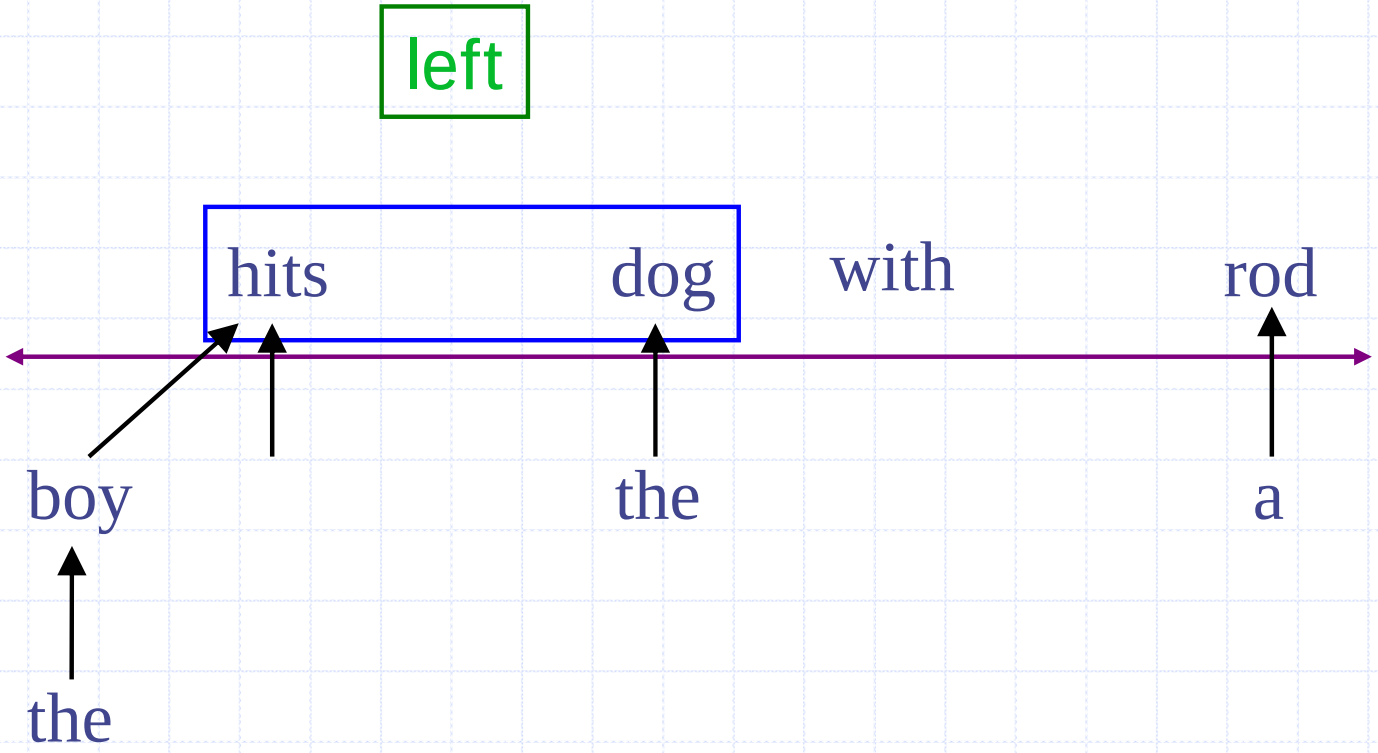
Yamada法による英語の係り受け解析の例



処理すべき単語対

参照文脈

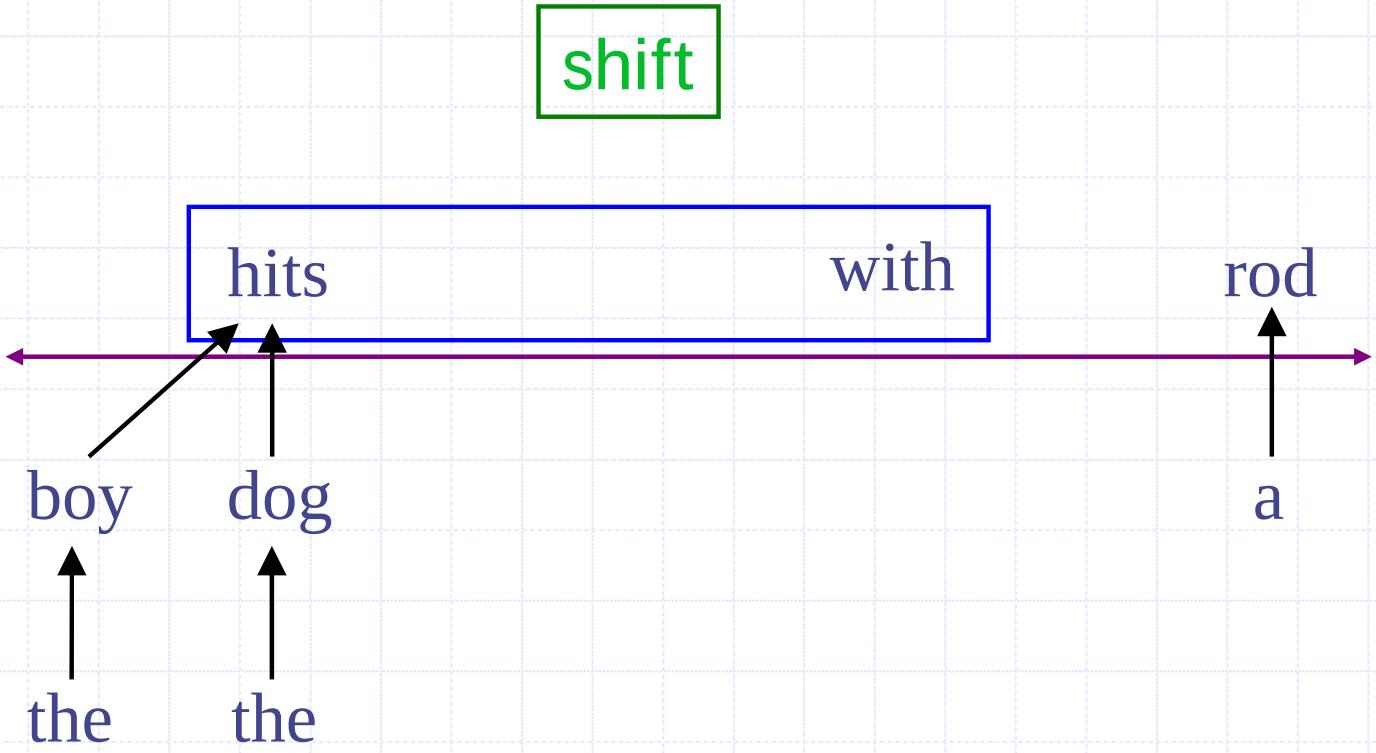
Yamada法による英語の係り受け解析の例



処理すべき単語対

参照文脈

Yamada法による英語の係り受け解析の例

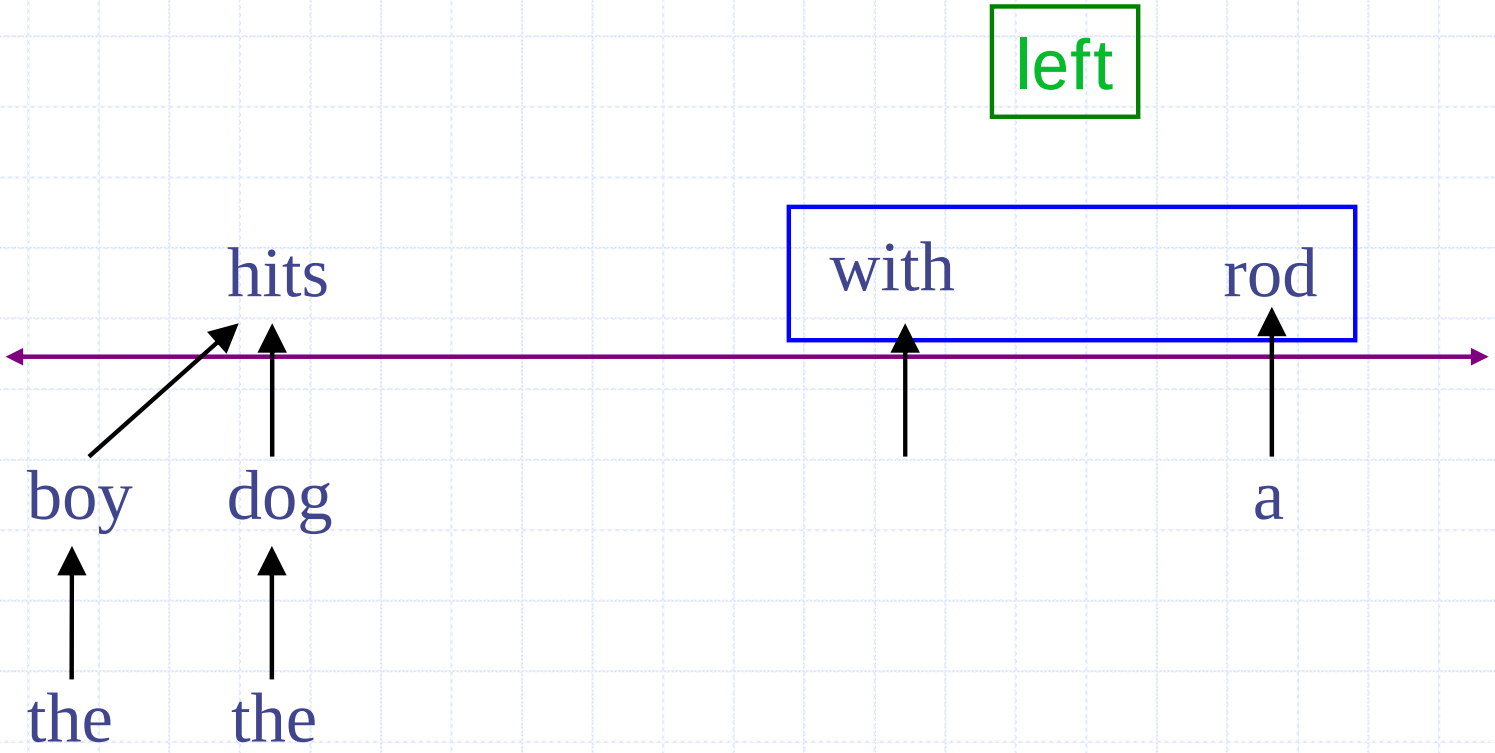


処理すべき単語対

参照文脈

NAIST

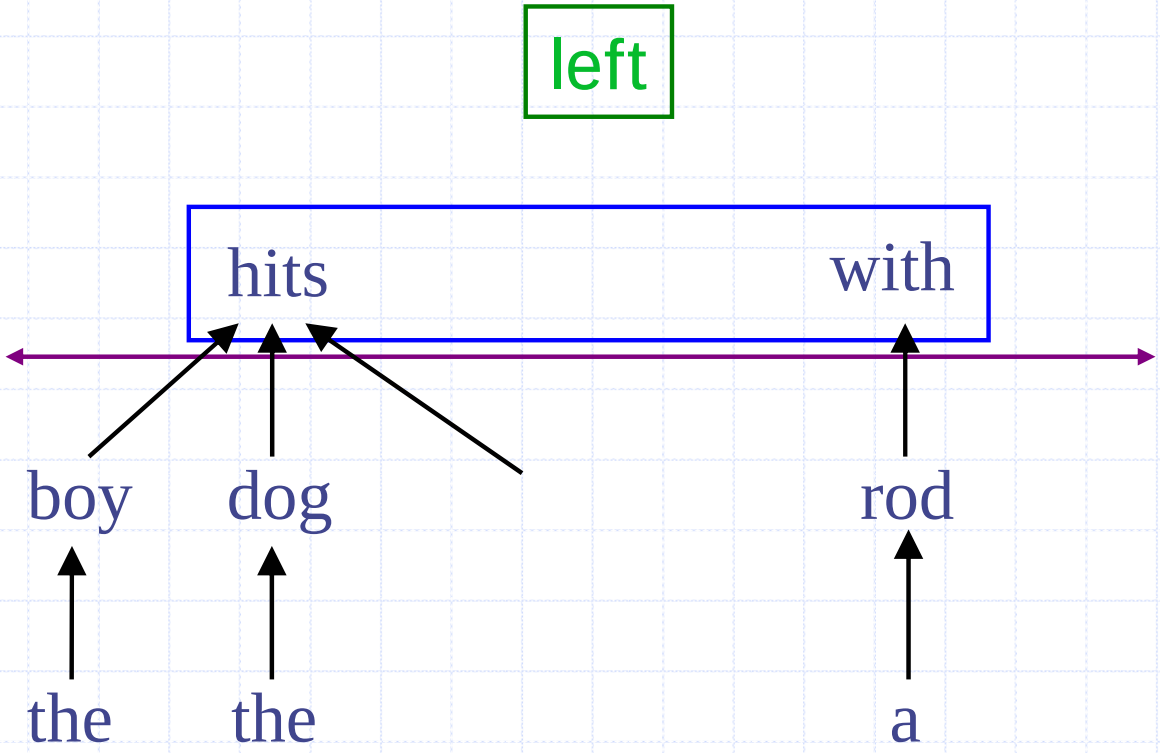
Yamada法による英語の係り受け解析の例



処理すべき単語対

参照文脈

Yamada法による英語の係り受け解析の例

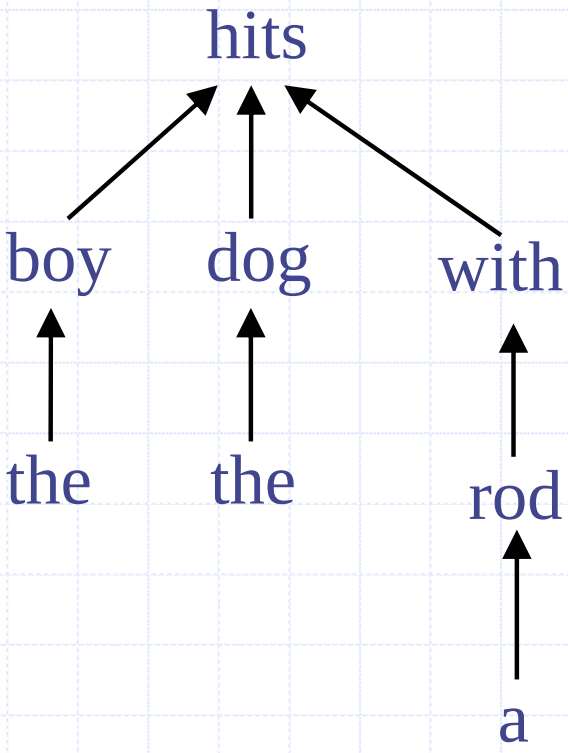


処理すべき単語対

参照文脈

Yamada法による英語の係り受け解析の例

処理の終了



統計的言語解析の貢献

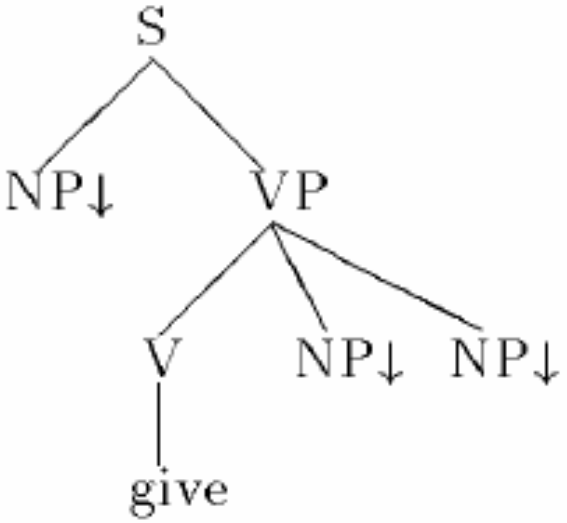
- 曖昧性の解消
  - 入力文に対して，単一（あるいは，順序付き）の解を返してくれる
- 頑健性の問題
  - 特に，依存構造（係り受け）解析は，どんな入力に対しても，ともかく解を返す

NAIST

制約に基づく文法の語彙化の流れ

- 個別の文法規則を持たず，ほとんどの文法情報を語がもつ
  - Lexicalized TAGでは，規則は2種類
  - HPSGでは，規則は4～5種類

Lexicalized TAG　　HPSG　　（それぞれgiveの記述例）



$$\begin{bmatrix} word & & \\ \text{VAL} & \begin{bmatrix} \text{SPR} & \langle \boxed{1} \rangle \\ \text{COMPS} & \langle \boxed{2}, \boxed{3} \rangle \end{bmatrix} & \\ \text{ARG-ST} & \langle \boxed{1}\,\text{NP}_i, \boxed{2}\,\text{NP}_j, \boxed{3}\,\text{NP}_k \rangle & \\ \text{SEM} & \begin{bmatrix} \text{INDEX} & s & \\ \text{RESTR} & \langle \begin{bmatrix} \text{RELN} & \text{give} \\ \text{SIT} & s \\ \text{GIVER} & i \\ \text{GIVEN} & j \\ \text{GIFT} & k \end{bmatrix} \rangle & \end{bmatrix} & \end{bmatrix}$$

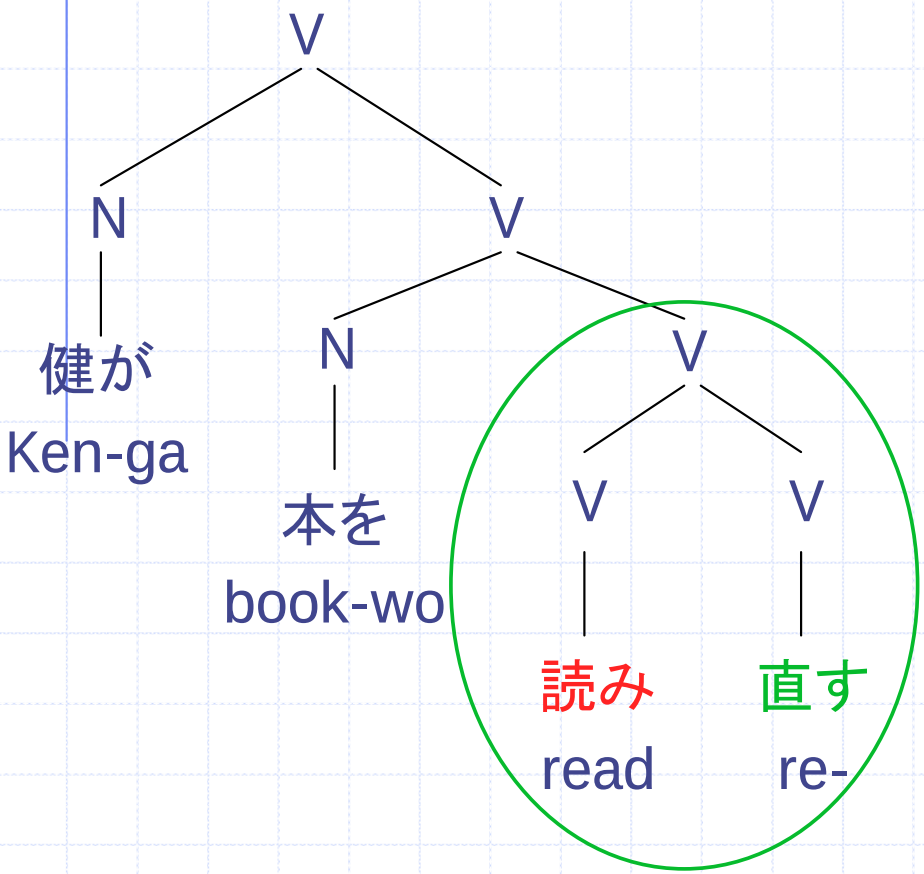
制約文法の利点

詳細な文法現象の記述

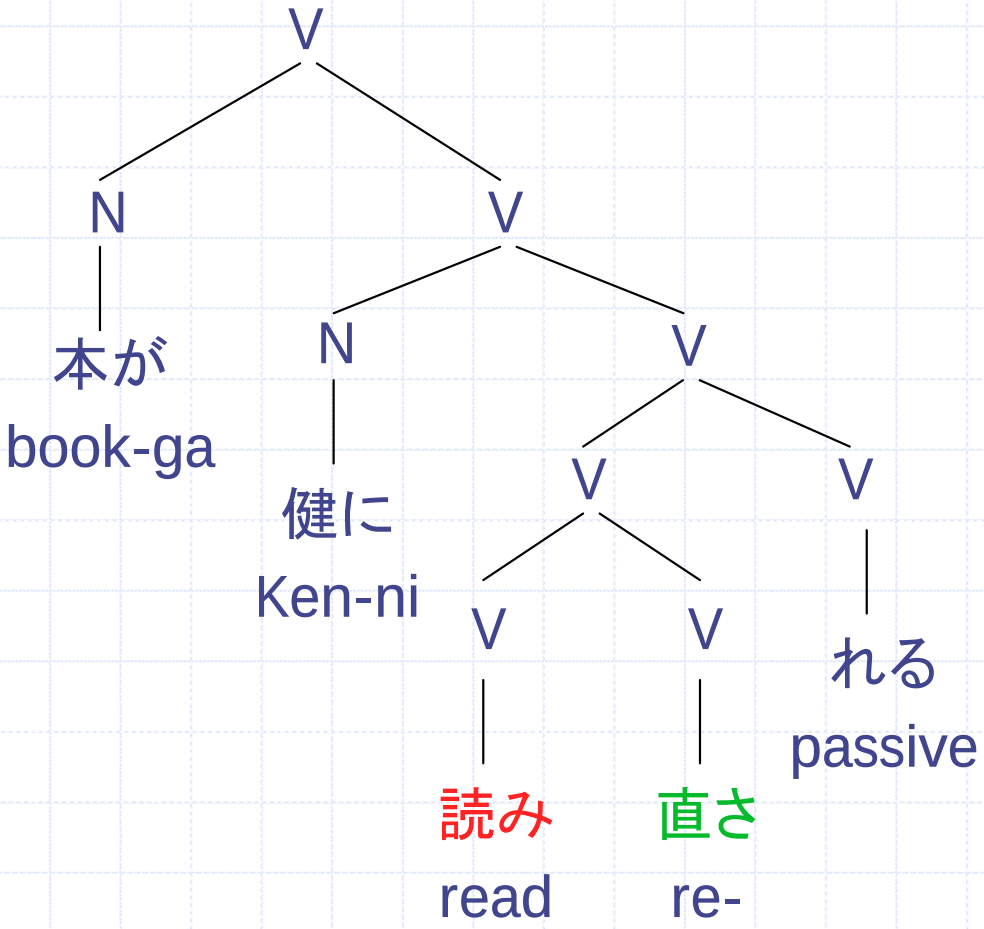
健が本を読み直した (Ken re-reads the book.)
本が健に読み直された (The book is re-read by Ken)
健が本を読みそびれた (Ken fails to read the book.)
*本が健に読みそびれられた (*The books fails to be read by Ken.)

HPSG による説明
「直す」は語彙的複合をつくる動詞
「そびれる」は統語的複合をつくる動詞
これらを各語の内部構造の記述によって区別できる

「直す」: 語彙的複合動詞

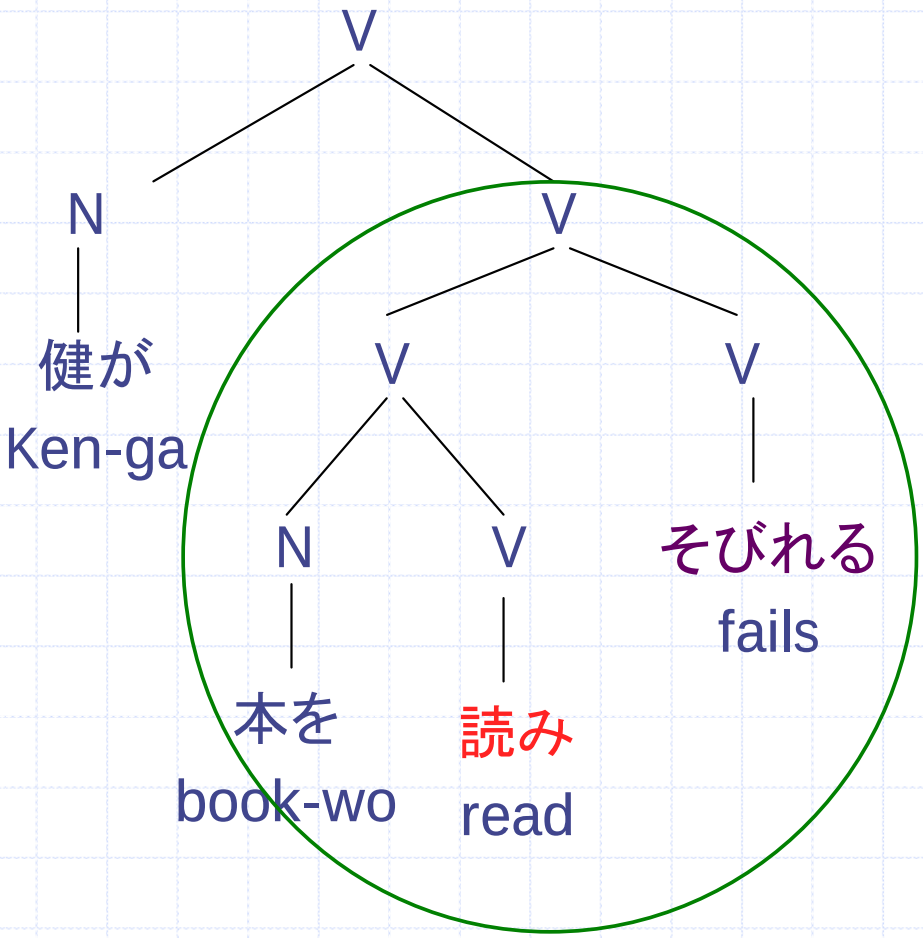


「直す」は単語（動詞）
を引数として取る



他動詞としての「読み直す」
が受動態を取り得る

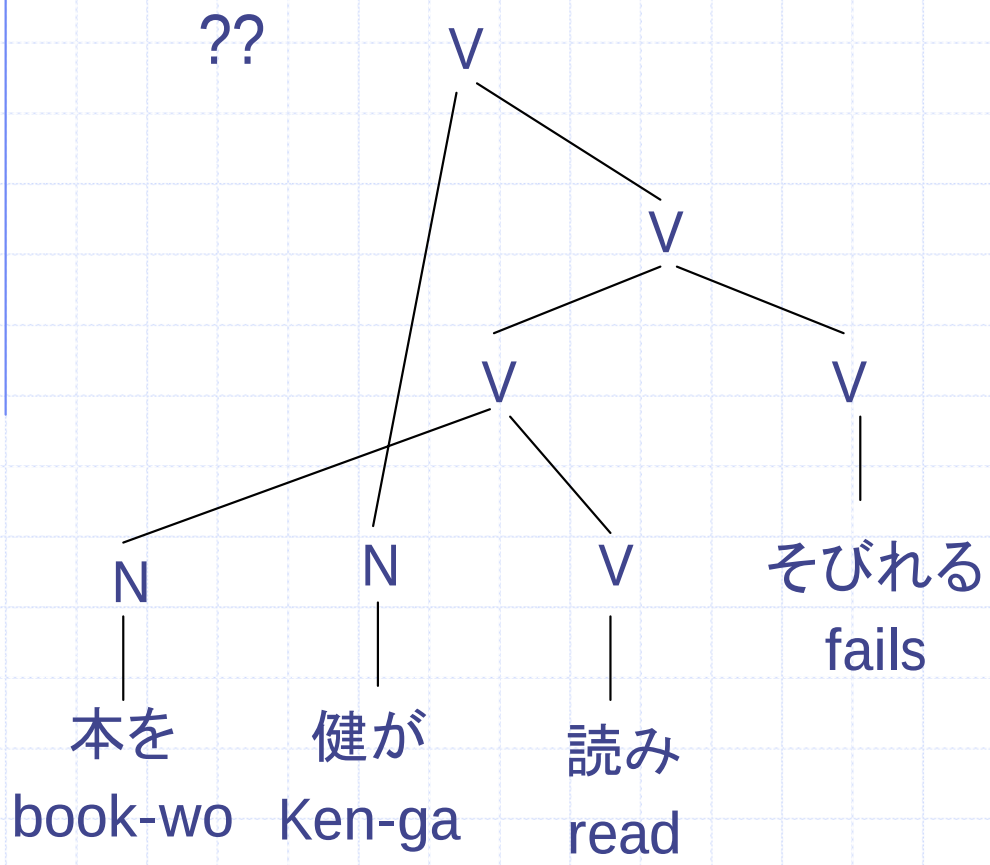
「そびれる」統語的複合動詞



「読みそびれる」は受動態不可

「そびれる」が目的語を既にもつ動詞句を引数として取るため

「そびれる」を含むかき混ぜ構文



この文は文法的に正しいが，句構造によって表現することができない

依存構造（係り受け）木を利用することで，この現象を回避できる

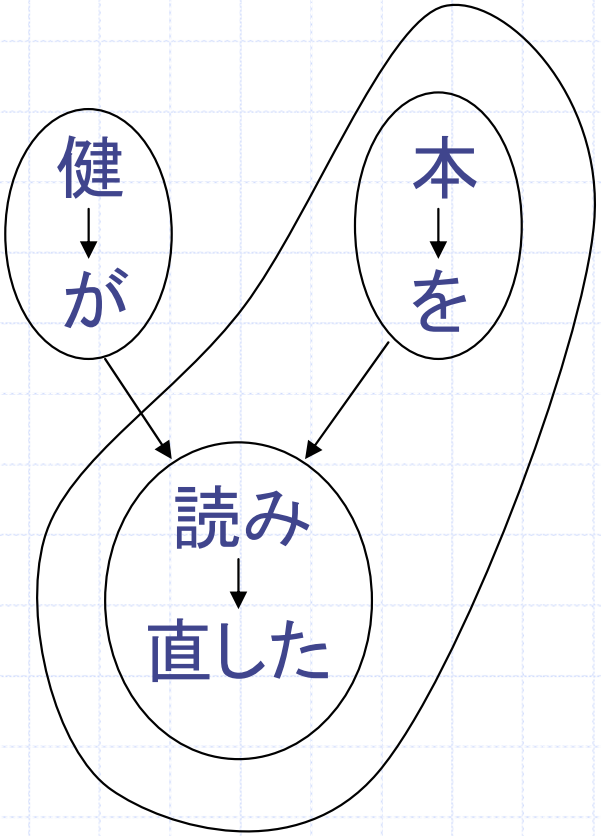
Use of Dependency as Control Information

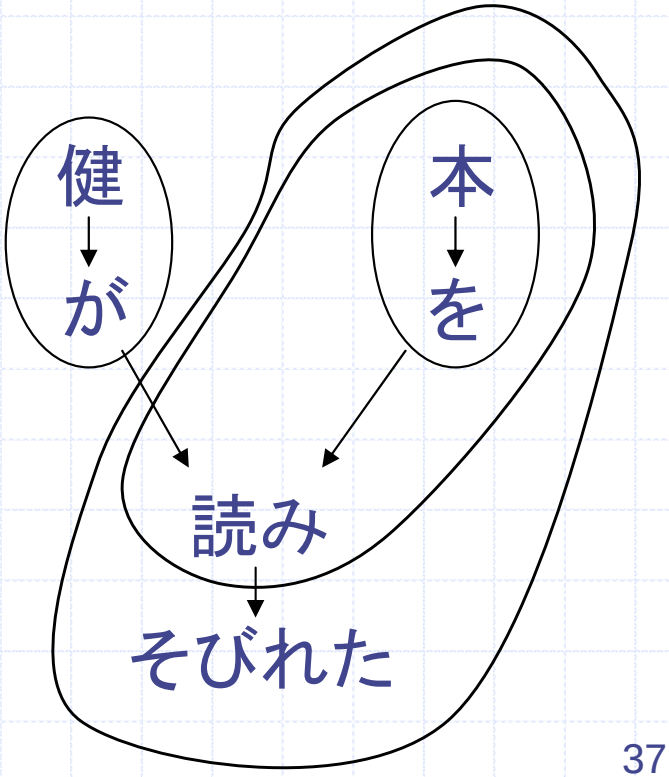
Non-projectivity (caused by scrambling is easily handled

健が	本を	読み直した
本を	健が	読み直した

健が	本を	読みそびれた
本を	健が	読みそびれた

book-wo read





近年の語彙意味論の進展の例

- Generative Lexicon [Pustejovsky 95]
  - 語の統語・意味構造を素性構造によって記述
  - 文法的不適格な文に対して，制約を緩和する3つの演算を定義 (強制：coersion，共構成：co-composition，選択束縛：selective binding)
  - 強制は：関数の引数のタイプを変更．共構成：関数のタイプを変更．選択束縛：引数の部分構造を選択的に引数とする
    - “He began the book.” は “begin” が事象を表す目的語を予測しているため，不適格
    - 強制(Coersion)により “book” の意味構造が「事象」に変更される．語の中にその語にまつわる事象に関する情報が記述されているためにこれが可能になる．
      - 語が持つQualia構造：構成役割，形式役割，目的役割，主体役割をもち，目的・主体役割がその語が目的語・主語としてどのような事象と関係するかを記述

GL representation of “begin”

$$
\left[\begin{array}{l}
\textbf{begin} \\
\text{ARGSTR} = \left[\begin{array}{l} \text{ARG1} = \mathbf{x} : \textbf{human} \\ \text{ARG2} = \mathbf{e_2} \end{array}\right] \\
\text{EVENTSTR} = \left[\begin{array}{l} \text{E}_1 = \mathbf{e_1} : \textbf{transition} \\ \text{E}_2 = \mathbf{e_2} : \textbf{transition} \\ \text{RESTR} = <_\propto \\ \text{HEAD} = \mathbf{e_1} \end{array}\right] \\
\text{QUALIA} = \left[\begin{array}{l} \text{FORMAL} = \mathbf{P}(\mathbf{e_2}, \mathbf{x}) \\ \text{AGENTIVE} = \textbf{begin\_act}(\mathbf{e_1}, \mathbf{x}, \mathbf{e_2}) \end{array}\right]
\end{array}\right]
$$

Qualia structure of a “book”

$$
\left[\begin{array}{l}
\textbf{book}(\mathbf{x}, \mathbf{y}) \\
\text{CONST} = \textbf{bound\_pages}(\mathbf{x}) \vee \textbf{disk}(\mathbf{x}) \\
\text{FORMAL} = \textbf{information}(\mathbf{y}) \\
\text{TELIC} = \textbf{read}(\mathbf{T}, \mathbf{w}, \mathbf{y}) \\
\text{AGENTIVE} = \textbf{artifact}(\mathbf{x}) \wedge \textbf{write}(\mathbf{T}, \mathbf{z}, \mathbf{y})
\end{array}\right]
$$

共構成の例

$$\left[\begin{array}{l}\textbf{bake}\\ \text{ARGSTR} = \left[\begin{array}{l}\text{ARG1} = \boxed{1}\left[\begin{array}{l}\textbf{animate}\\ \text{FORMAL} = \textbf{physobj}\end{array}\right]\\ \text{ARG2} = \boxed{2}\left[\begin{array}{l}\textbf{mass}\\ \text{FORMAL} = \textbf{physobj}\end{array}\right]\end{array}\right]\\ \text{EVENTSTR} = \left[\begin{array}{l}\text{E}_1 = \text{e}_1 : \textbf{process}\\ \text{HEAD} = \text{e}_1\end{array}\right]\\ \text{QUALIA} = \left[\begin{array}{l}\textbf{state\_change}\\ \text{AGENTIVE} = \textbf{bake\_act}(\text{e}_1, \boxed{1}, \boxed{2})\end{array}\right]\end{array}\right]$$

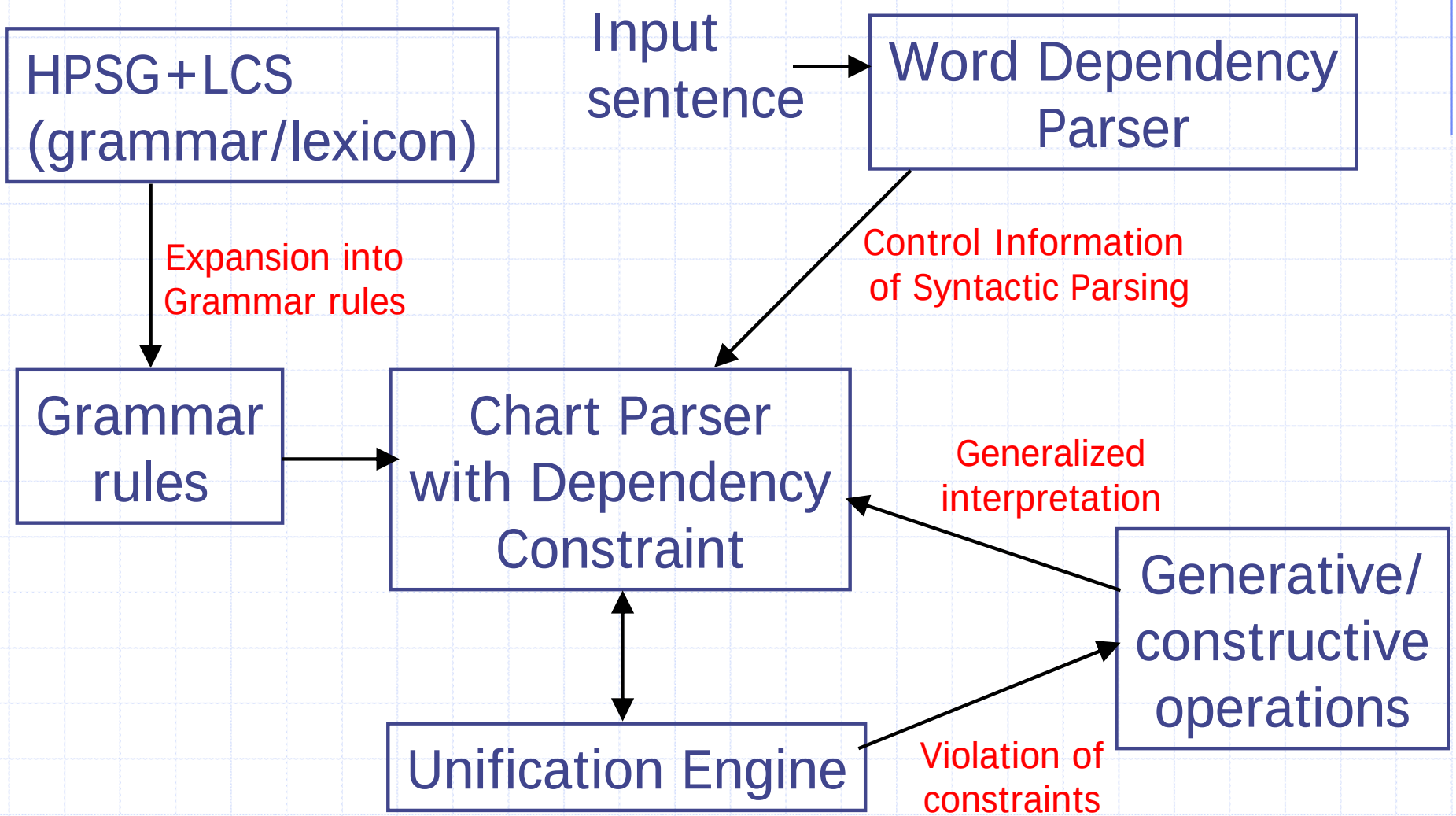
$$\left[\begin{array}{l}\textbf{cake}\\ \text{ARGSTR} = \left[\begin{array}{l}\text{ARG1} = \text{x} : \textbf{food}\\ \text{D\_ARG1} = \text{y} : \textbf{mass}\end{array}\right]\\ \text{QUALIA} = \left[\begin{array}{l}\text{CONST} = \text{y}\\ \text{FORMAL} = \text{x}\\ \text{TELIC} = \textbf{eat}(\text{e}, \text{z}, \text{x})\\ \text{AGENTIVE} = \textbf{artifact}(\text{x}) \wedge \textbf{bake\_act}(\text{e}', \text{w}, \text{y})\end{array}\right]\end{array}\right]$$

- bake と cakeの記述から,
- I bake a cake は正しい文ではないことになる.
- bakeは「材料」を目的語に予測しており, cakeは人工物のため, 材料ではない, (bake a potato ならOK)
- 共構成によりbakeが単なるprocess動詞ではなく, 結果をもつcreation動詞に変更されることで, この文が解釈できる

処理および文法記述における語彙化の流れ

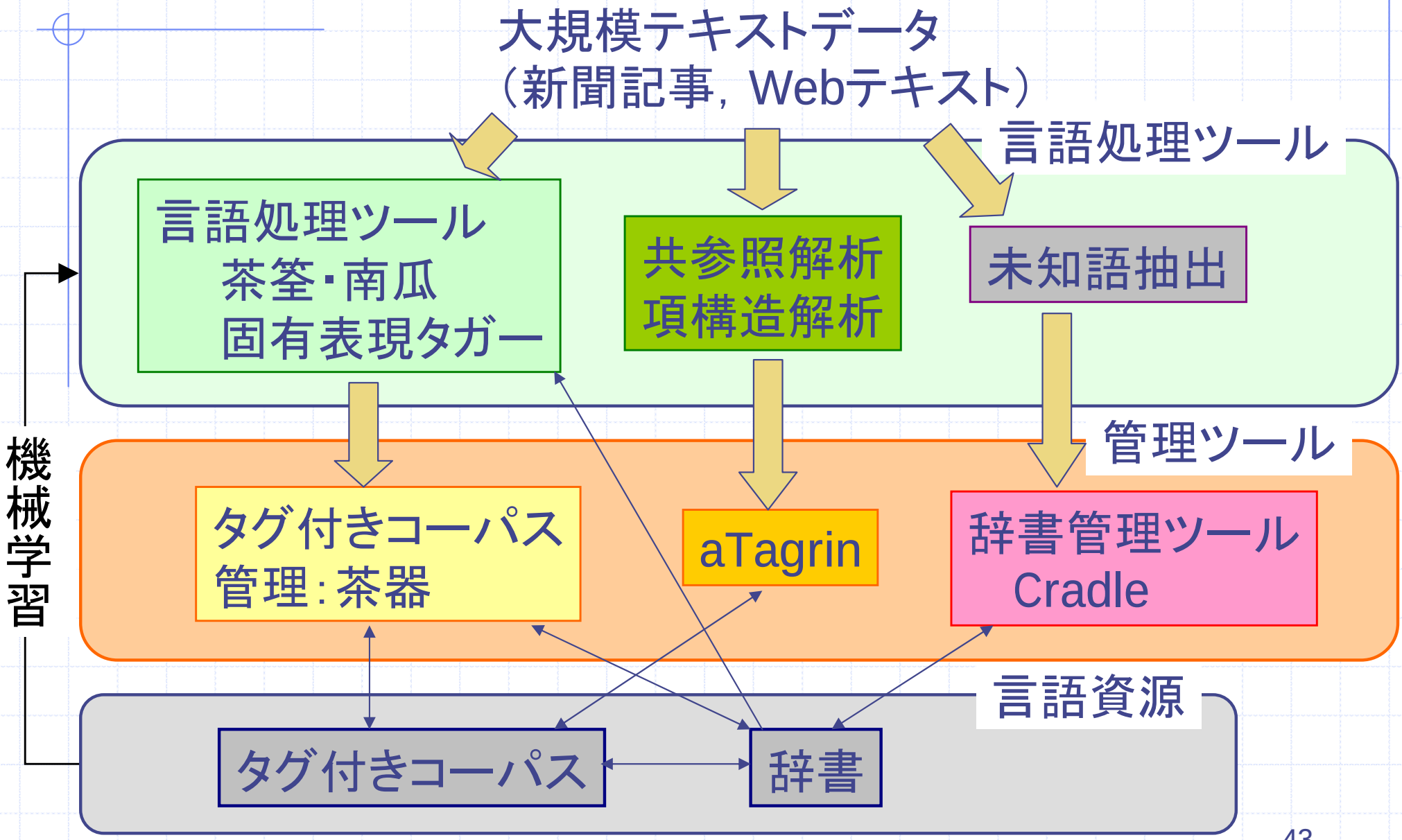
- 統計的言語解析
  - 単語の情報を素性として用いることにより，精度の高い言語解析が可能になってきた
  - 依存構造については，ハードな制約は存在しない
- 生成語彙
  - 意味表現と統語の干渉：統語制約を意味情報によりoverwrite
  - 動的な制約緩和を実現

Overview of the Integrated System



我々の研究グループの最近の活動：言語処理ツールとコーパス管理システム

NAIST



NAIST

我々のグループで構築し公開している言語処理ツール

- NLP Tools Based on Machine Learning
  - Japanese Morphological Analyser:
    - ChaSen [Asahara 00] – variable memory length HMM
      - Multi-lingual version: Japanese, Chinese, English
    - MeCab [Kudo 04] – Conditional Random Fields
  - Japanese Dependency Parser: CaboCha [Kudo 02]
  - English and Chinese Word Dependency Parsers [Yamada 03, Chen 04]
  - General Purpose Chunker: YamCha [Kudo 01]
    - Named Entity Recognition [Asahara 03] [Watanabe 07]
    - Unknown Word Identifier: bar [Asahara 04]
  - Anaphora Resolution and Co-reference Analysis
    - Japanese zero-pronoun and co-reference [Iida 03, 05, 06]
- Management Tools for Linguistic Data
  - Annotated Corpus Management Tool: ChaKi [Matsumoto 06]
  - Dictionary Management Tool: Cradle
  - General Purpose Annotation Tool: aTagrin

まとめ

- 制約に基づく文法
  - 曖昧性の問題：唯一解の選択，順序付け
  - 頑健性の問題：例外事象への対応
- 語彙情報を中心にした言語解析
  - 統計的言語解析：単語（文節）係り受け
  - 語彙意味と生成的演算に基づく制約の動的緩和
  - これらの融合
- 応用および今後
  - Webからの評判・意見情報マイニング
  - 日本語国家コーパスプロジェクト
  - 言語解析手法の精緻化
    - 全域情報を用いた全域最適化