共生社会特論

第2回
ルールベース機械翻訳

小川 泰弘

講義用ページ

http://www.kl.itc.nagoya-u.ac.jp/~yasuhiro/MT/

翻訳 (Translation)

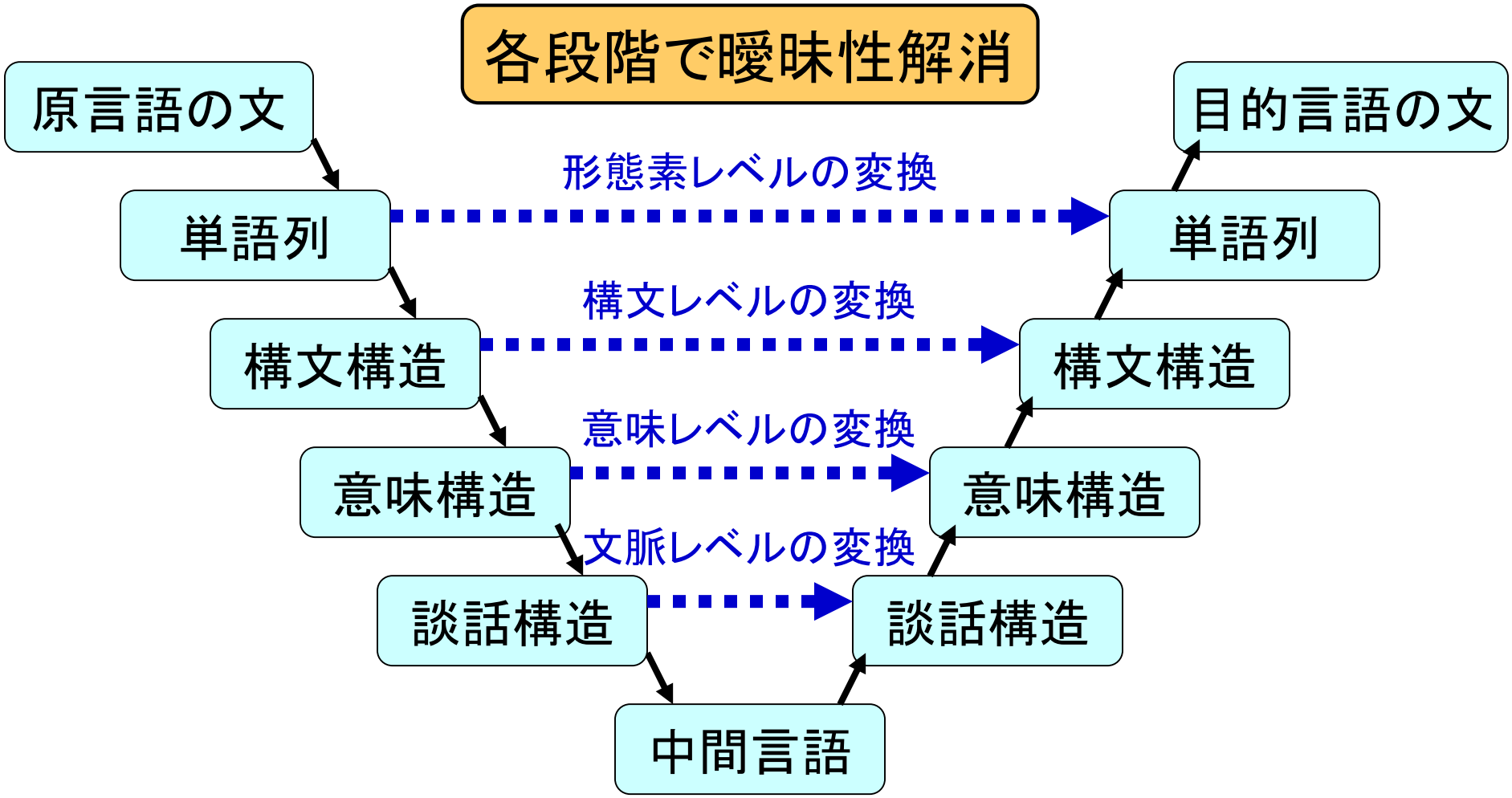
- 原言語 (source language)から
  目的言語 (target language)への変換
  - 翻訳
  - 通訳 (interpretation)
  - 翻字 (transliteration)
  - プログラムのコンパイル

機械翻訳

機械翻訳 (Machine Translation)

- 単語直接方式 (direct method)
  - 単語を直訳するだけ
- 変換方式 (transfer method)
  - 入力文を解析し、ある段階で目的言語の構造へ変換する
- 中間言語方式 (pivot method)
  - 複数の言語間の翻訳に中間言語を用意
    - 実在の言語（英語）
    - 概念レベル

機械翻訳における処理レベル



アプローチ

- ルールベース手法
  - 文法知識の規則化
  - 全ルールの記述は簡単ではない
    - 多数の例外
- 統計的手法
  - データから確率的に学習
  - 文法的にありえない解析をする場合も

形態素解析

形態素解析 (Morphological Analysis)

文を形態素に分割する作業

- 分割･トークン化
- 語彙化
- 品詞タグ付与

辞書が必要

分割・トークン化 (tokenization)

一連の文字列を意味のある塊（トークン）へ

- 空白で区切る

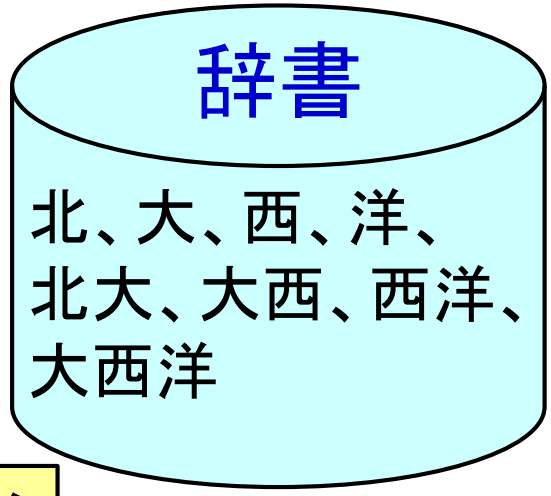
  ⇒以下の例はどうする？

- *data-base*
- *cat's eye*
- *$1,005.98*
- 日本語・中国語・タイ語
- ドイツ語の複合語
  Lebensversicherungsgesellschaft

Leben#Versicherung#Gesellschaft

分割の曖昧性

- 北大西洋 ⇨ 北 大西洋
  - ✗ 北 大西 洋
  - ✗ 北大 西洋
  - ✗ 北 大 西 洋
- 辻元議員 ⇨ 辻元 議員
  - 辻 元 議員



？

アルゴリズム

- 最長一致法　北大　西洋
- 分割数最小法　北大　西洋　北　大西洋
- 字種区切り法　今日はマウンテンへ行く
- 接続コスト最小法

現在の主流は

統計的な接続コスト最小法

短単位と長単位

- 短単位

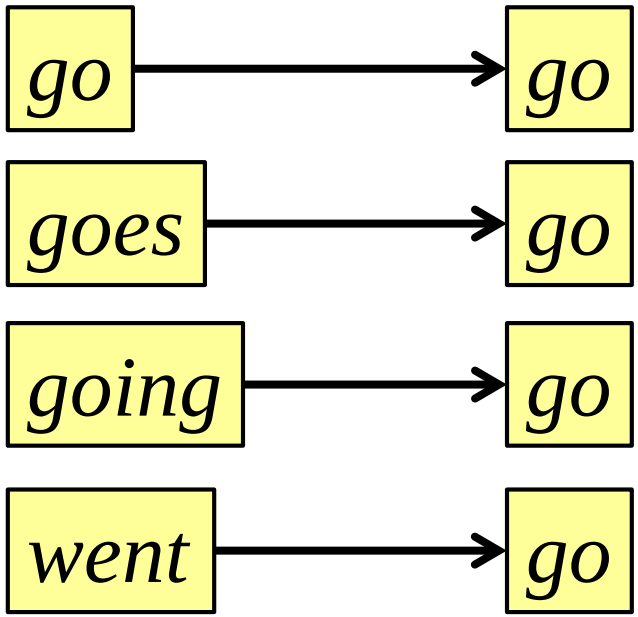
生命	保険	会社	に	つい	て	お	話し	いたし	ます

- 長単位

生命保険会社	について	お話しいたし	ます

語彙化 (lemmatization)

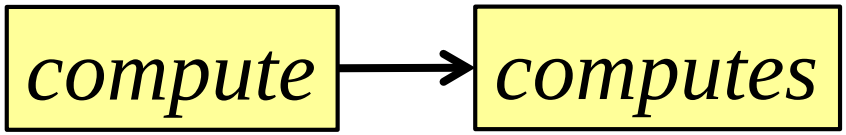
- 語形変化を処理し原形にする



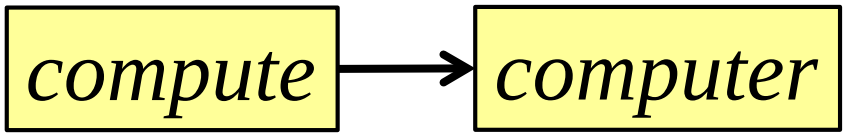
- 語幹化(stemming)と共通する点も多い
  - 語幹化は品詞を認定しない

屈折と派生

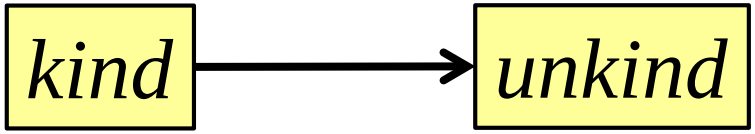
- 屈折 (inflection)（活用: conjugation）
  - 品詞は変化しない
  - 文法的素性（単数・複数、過去・現在）を示す



- 派生 (derivation)
  - 品詞が変化することもある



  - 意味が変わることもある



ステマー (stemmer)

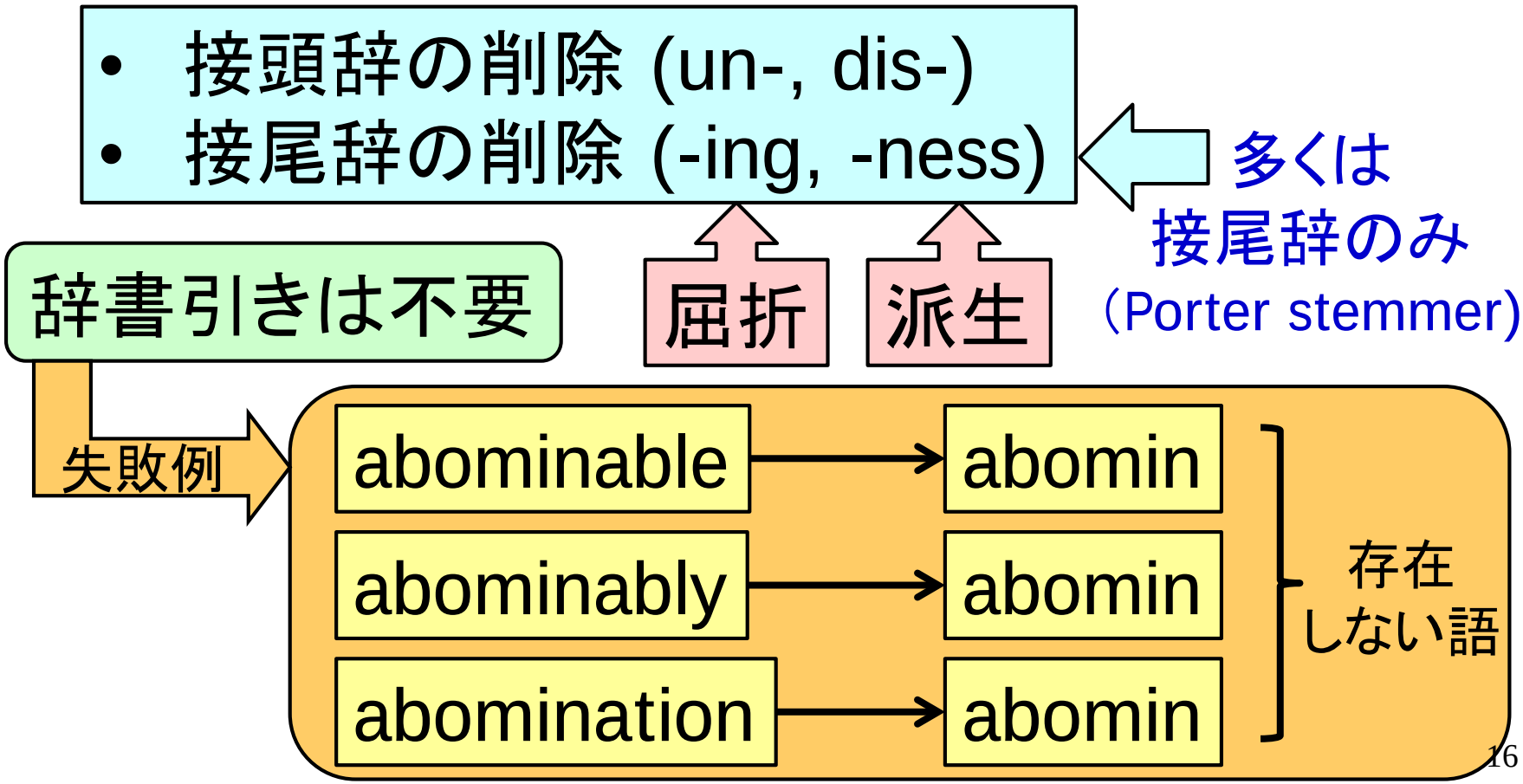
- 語幹化器
  - ステマーによっては派生も処理する

- 品詞タグ付与と併用 → 屈折のみ処理
- 不規則動詞は辞書が必須
- 曖昧な例 *lay*
  - *lay* の原形
  - *lie* の過去形

ヒューリスティック・ステマー

正確な解析は必ずしも必要ない

- 文法的な規則より経験則による正規化



日本語：動詞の活用処理

- 動詞や助動詞を終止形に

書きました → 書く ます た

させました → する ます た

- 曖昧性がある場合

行った → 行く た　　きた → きる た

行った → 行う た　　きた → くる た

アルゴリズム

- 活用形展開方式

  書か　書き　書く　書け　書こ　書い

- 活用語尾分離方式

  書　か　き　く　け　こ　い

- 形態素解析器の内部処理

品詞タグ付与 (POS tagging)

語に品詞タグを付与

- 名詞 (Noun)
- 動詞 (Verb)
- 形容詞 (ADJective)
- 副詞 (ADVerb)
- 助動詞 (AUXiliary verb)
- 前置詞 (Preposition)
- 限定詞 (DETerminer)　冠詞や those, my

曖昧な例

Visiting aunts can be a nuisance.

ADJ | N-Pl | AUX | V-inf-be | DET-Indef | N-sg

Visiting aunts can be a nuisance.

V-Prog | N-Pl | AUX | V-inf-be | DET-Indef | N-sg

いずれが正しいかは文脈などに依存

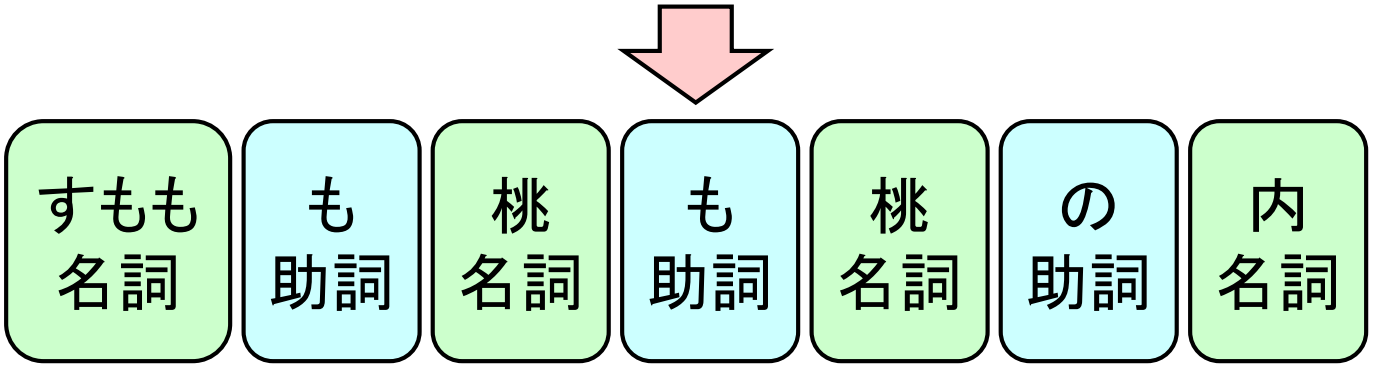
有名な例

Time flies like an arrow.

- 光陰矢の如し　　（動詞は *flies*）
- 時蠅は矢を好む　　（動詞は *like*）
- 矢の速度を測るように、蠅の速度を測れ
  （動詞は *time*）

形態素解析の応用：かな漢字変換

すももももももものうち



きしゃのきしゃがきしゃできしゃした。

貴社の記者が汽車で帰社した。

検索エンジンへの応用

入力語が「日本」

- 形態素解析なしの場合
  - 昨日本を買った が発見されてしまう
- 形態素解析有の場合
  - 日本語の勉強 が発見されない場合がある

（「日本語」が1語として辞書にある場合）

Nグラム (N-gram)

- N個のまとまりを指す
- 文脈情報の一部として利用
- 形態素解析の代わりに使用されることも

文字Nグラム

北大西洋

- unigram (1-gram)

  北 大 西 洋

- bigram (2-ram)

  北大 大西 西洋

- trigram (3-gram)

  北大西 大西洋

単語Nグラム

Time flies like an arrow.

- unigram

Time *flies* *like* *an* *arrow*

- bigram

Time flies *flies like* *like an* *an arrow.*

品詞Nグラム

統計的に以下を決定する際に利用

- 品詞タグ付けの確率

N + N

N + V

出現回数を比較

- 形態素解析の接続コスト

接続可能行列

右連接属性 ＼ 左連接属性	名詞	動詞	形容詞	格助詞	活用語尾	名詞接尾辞	接頭辞	句読点
名詞語幹	○	○	○	○	－	○	○	○
動詞語幹	－	－	－	－	○	－	－	－
形容詞語幹	○	○	○	－	－	－	－	○
格助詞	○	○	○	－	－	－	○	○
名詞接尾辞	○	○	○	○	－	－	○	○
連体形接尾辞	○	○	○	－	－	－	○	○

接続コスト表

右連接属性 ＼ 左連接属性	名詞	動詞	形容詞	格助詞	活用語尾	名詞接尾辞	接頭辞	句読点
名詞語幹	20	30	30	5	–	5	30	20
動詞語幹	–	–	–	–	5	–	–	–
形容詞語幹	40	40	40	–	–	–	–	30
格助詞	10	10	10	–	–	–	10	40
名詞接尾辞	15	10	10	20	–	–	20	20
連体形接尾辞	10	40	50	–	–	–	10	50

接続コストの和が最小となる組合せを解とする

辞書

- 形態素解析に必須
  - 基本形
  - 活用
  - 品詞
  - その他（読み・意味）

辞書の構築

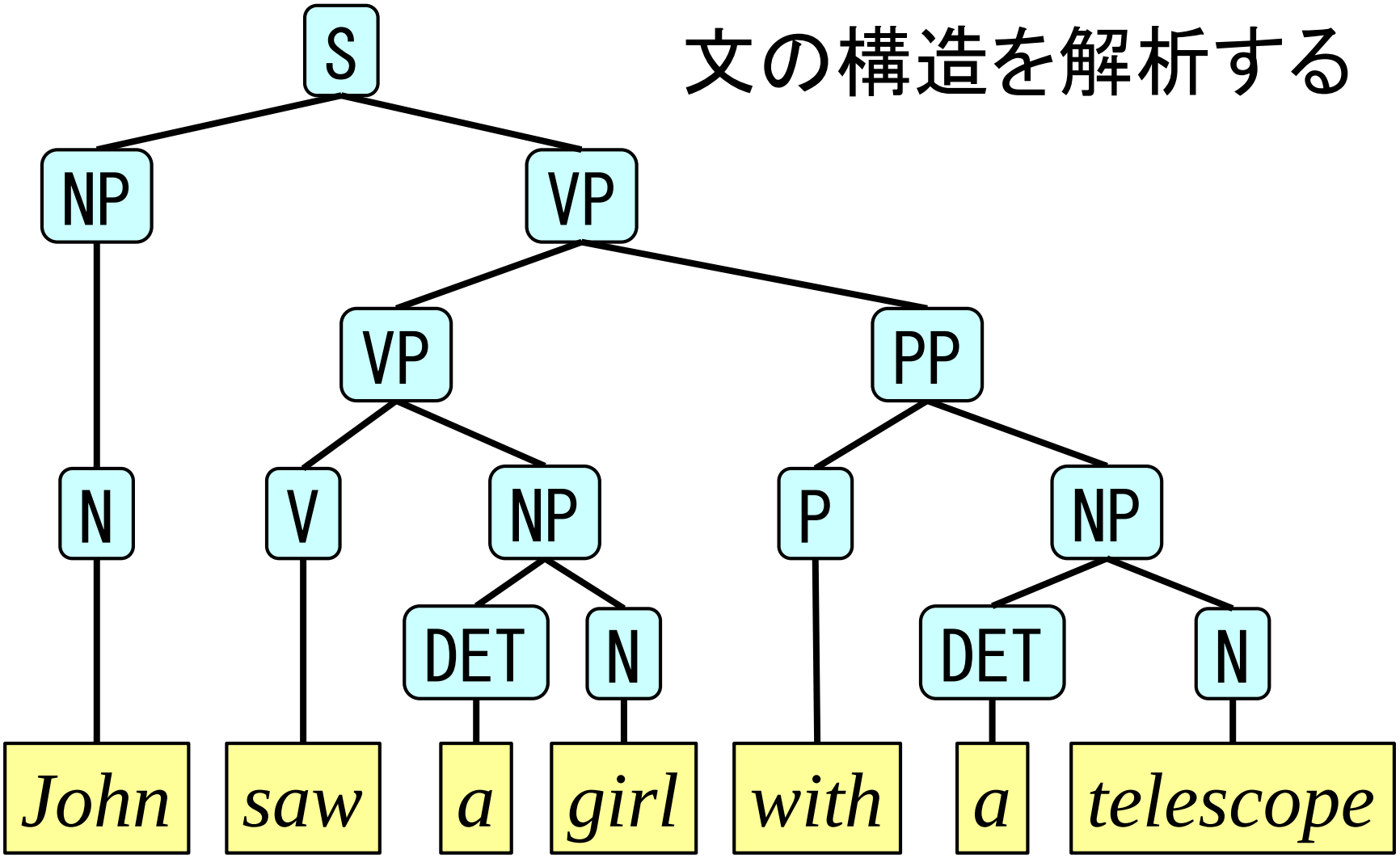
- 人手による作業
  - 時間がかかる
- ウェブからの収集
  - 膨大な量を短時間で構築
  - 専門用語・新語に対応

構文解析

句構造文法

構文解析 (Parsing)

文の構造を解析する



句構造文法 (Phrase Structure Grammar)

文脈自由文法 (CGF)

G = (V, T, S, P)

V: 非終端記号(変数)の集合

N: 終端記号(単語)の集合

S: 開始記号

P: 生成規則の集合

- 文を文法Gで生成する
- 文を文法G(に基づく構文解析器)で受理する

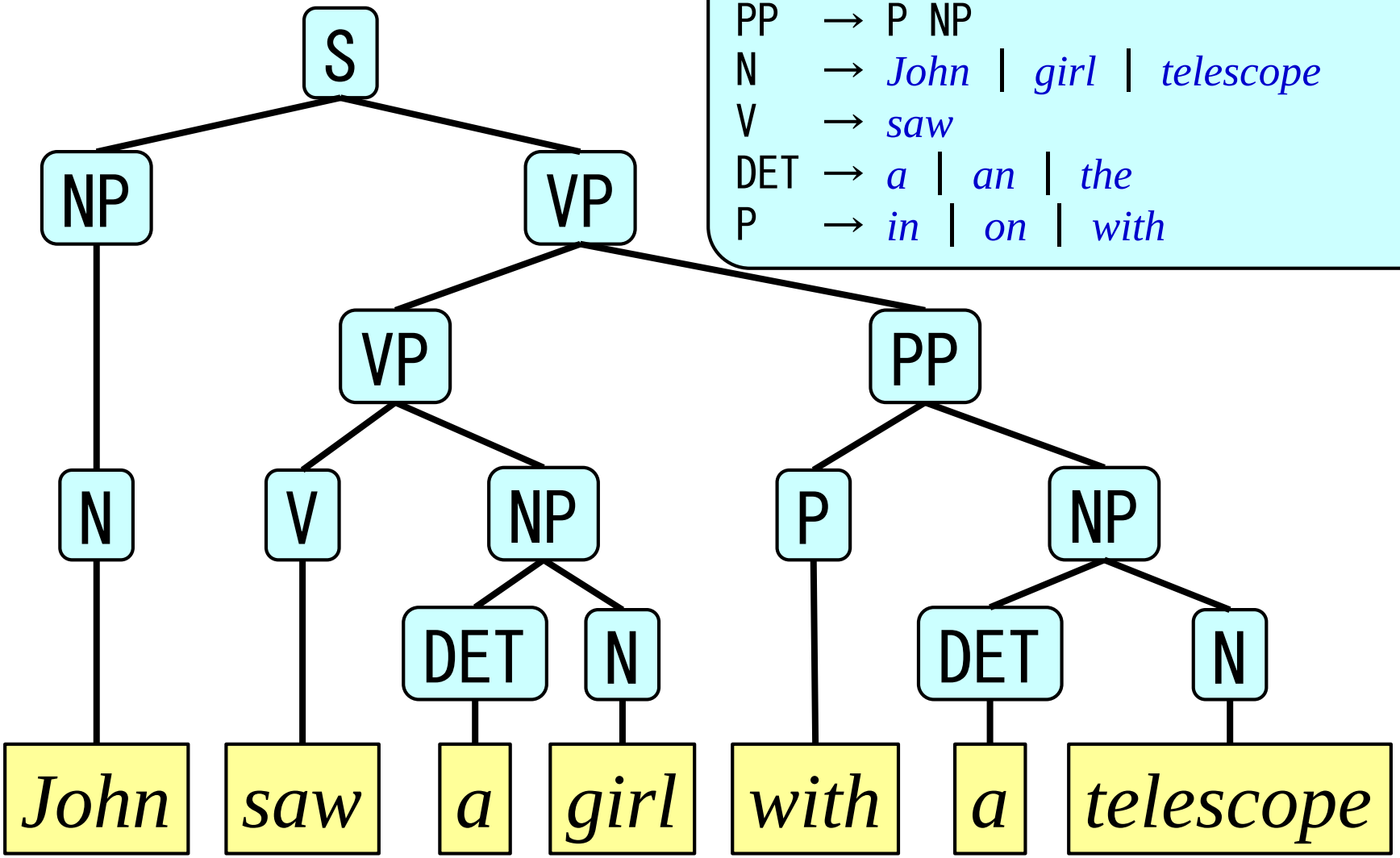
文脈自由文法の例

G = (V, T, S, P)

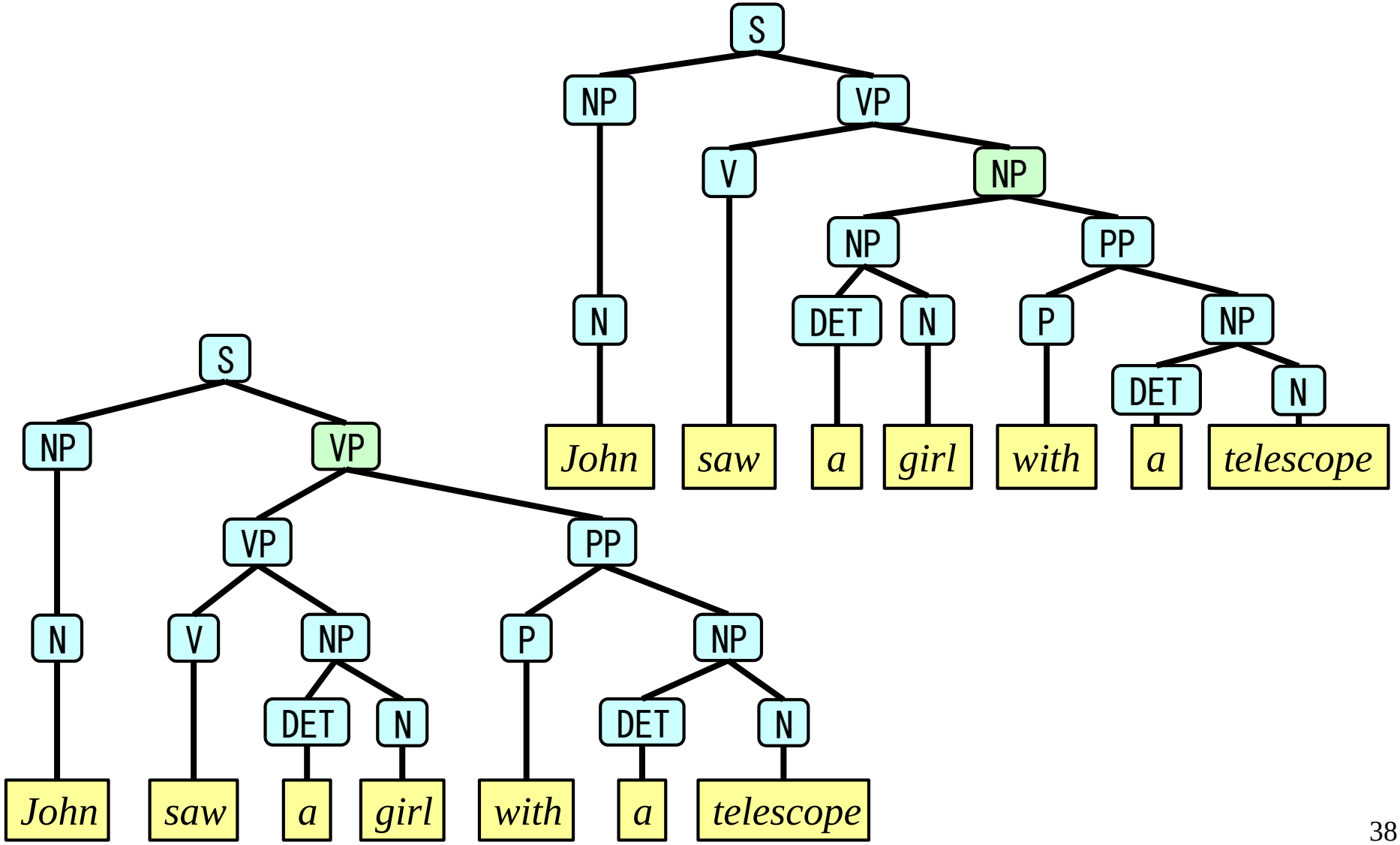
S → NP VP
NP → N | DET N | ADJ N | NP PP
VP → V | V NP | VP PP
PP → P NP
N → *John* | *girl* | *telescope*
V → *saw*
DET → *a* | *an* | *the*
P → *in* | *on* | *with*

構文木(syntax tr

```
S   → NP VP
NP  → N | DET N | ADJ N | NP PP
VP  → V | V NP | VP PP
PP  → P NP
N   → John | girl | telescope
V   → saw
DET → a | an | the
P   → in | on | with
```



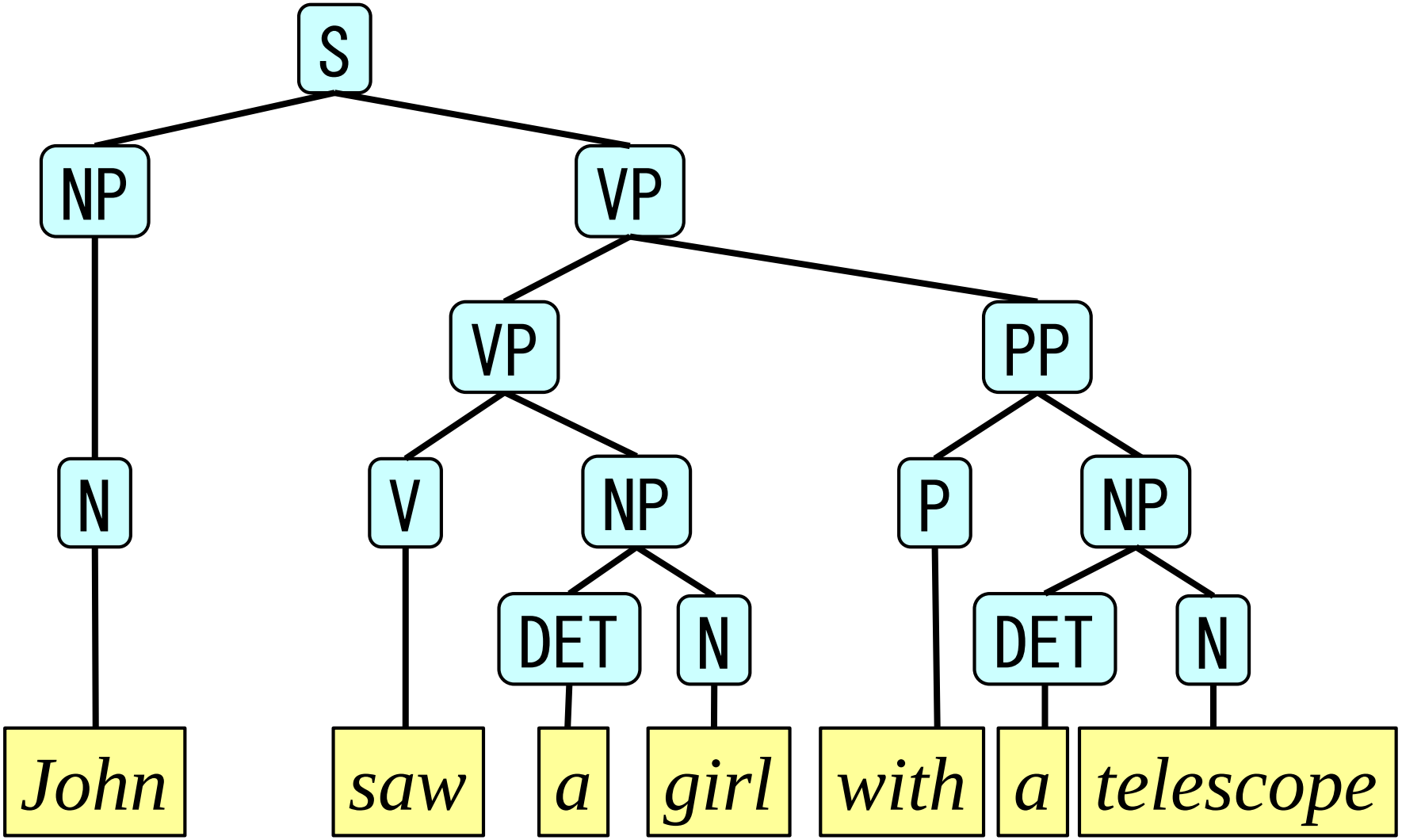
曖昧な構文木


S
NP
VP
V
NP
NP
PP
N
DET
N
P
NP
DET
N
John
saw
a
girl
with
a
telescope
S
NP
VP
VP
PP
N
V
NP
P
NP
DET
N
DET
N
John
saw
a
girl
with
a
telescope

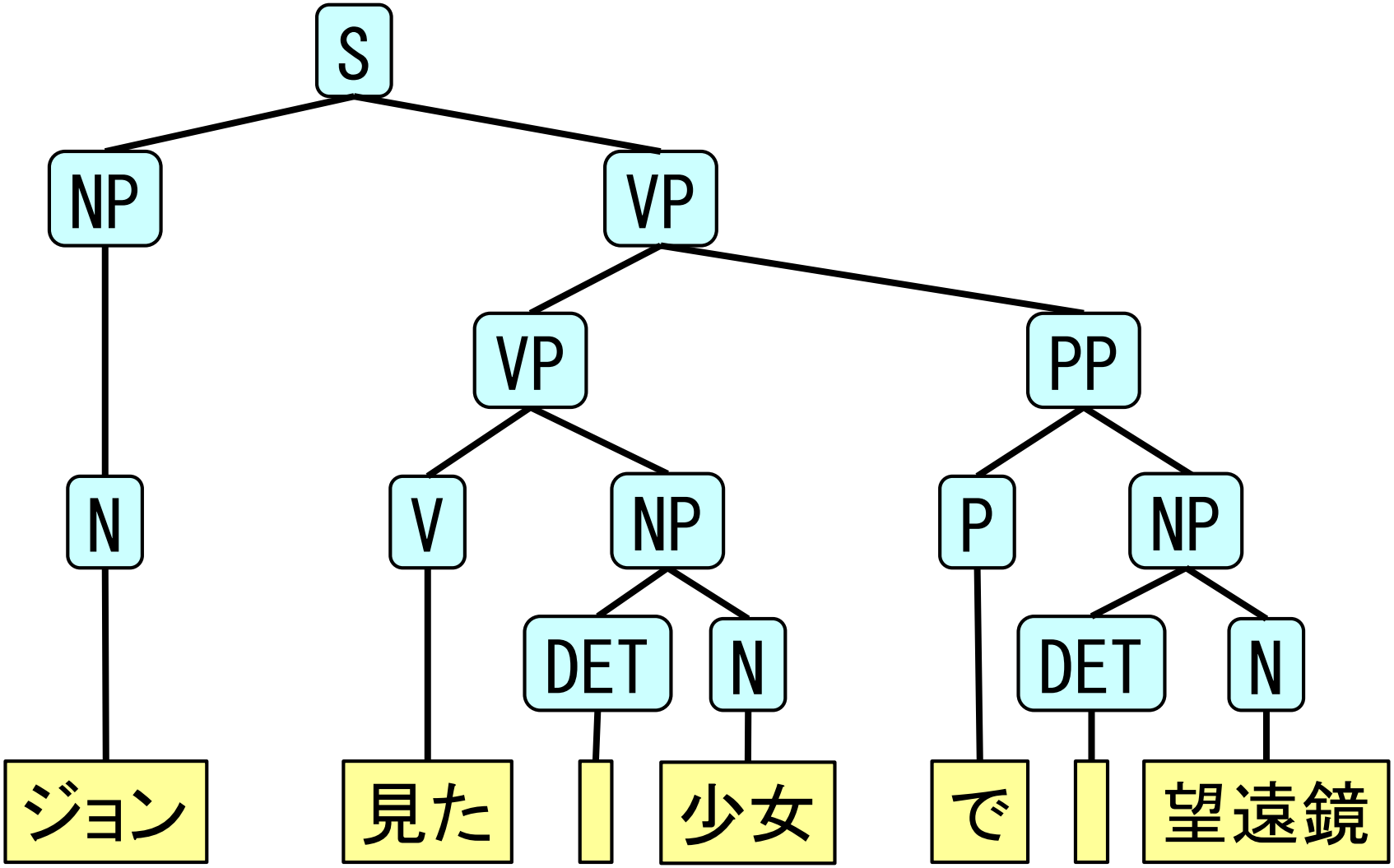
構文解析

- CYK法
  - チョムスキー標準形にのみ適用可能
- チャート法
- LR法
  - LR文法にのみ適用可能
  - コンパイラなどで使用
- LL法
  - LL文法にのみ適用可能
  - コンパイラなどで使用

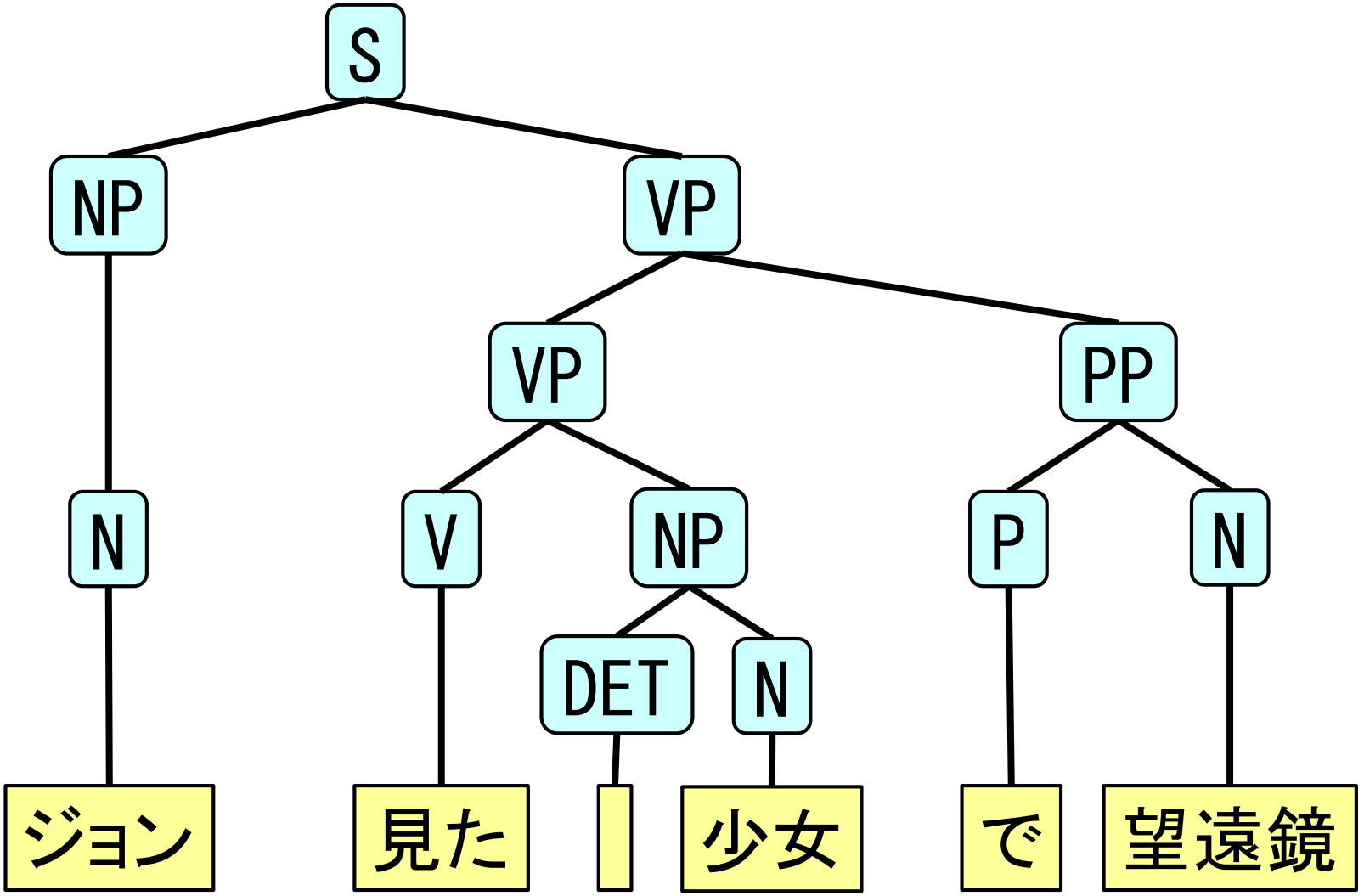
構文木の変換による翻訳



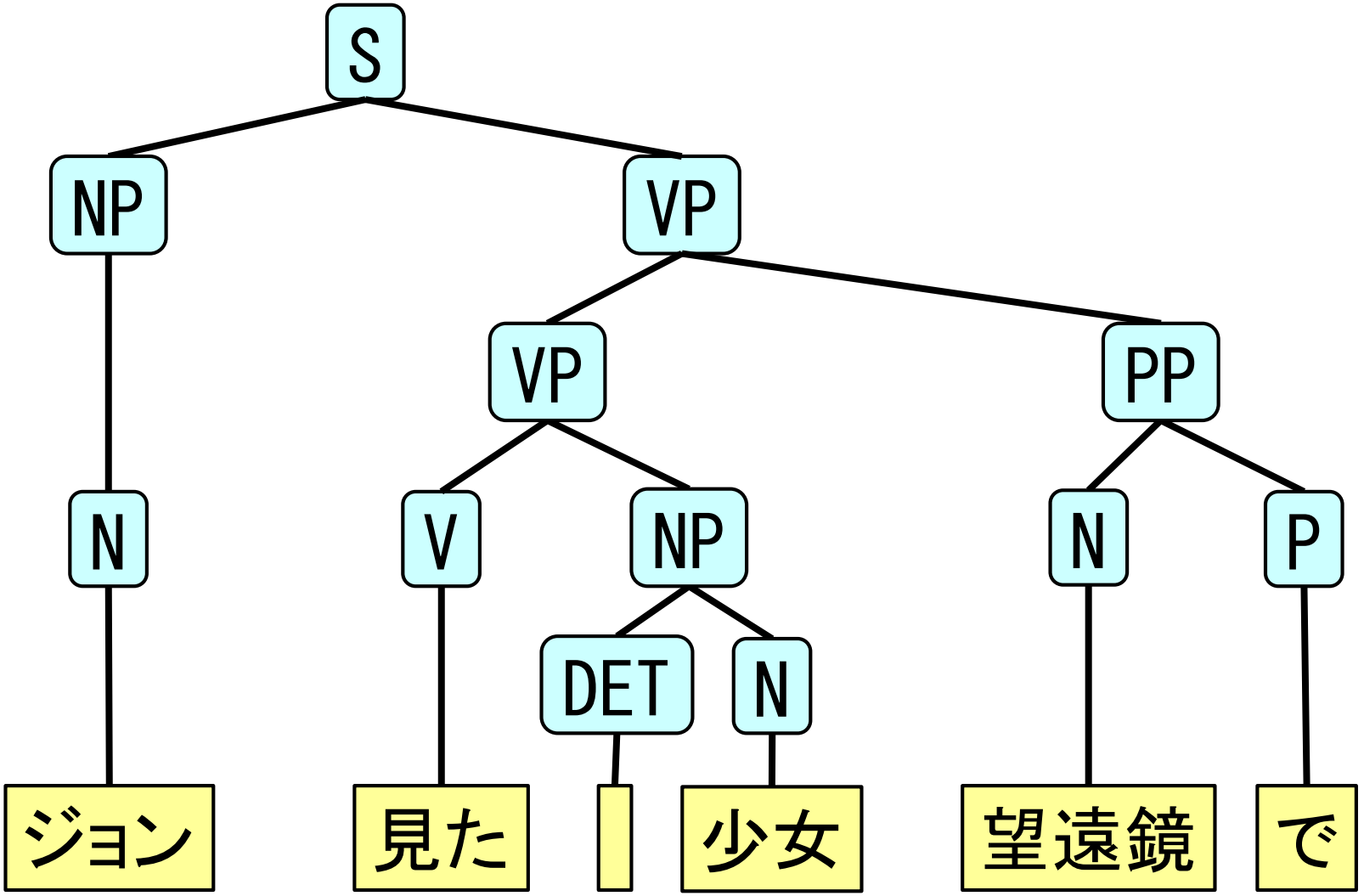
構文木の変換による翻訳



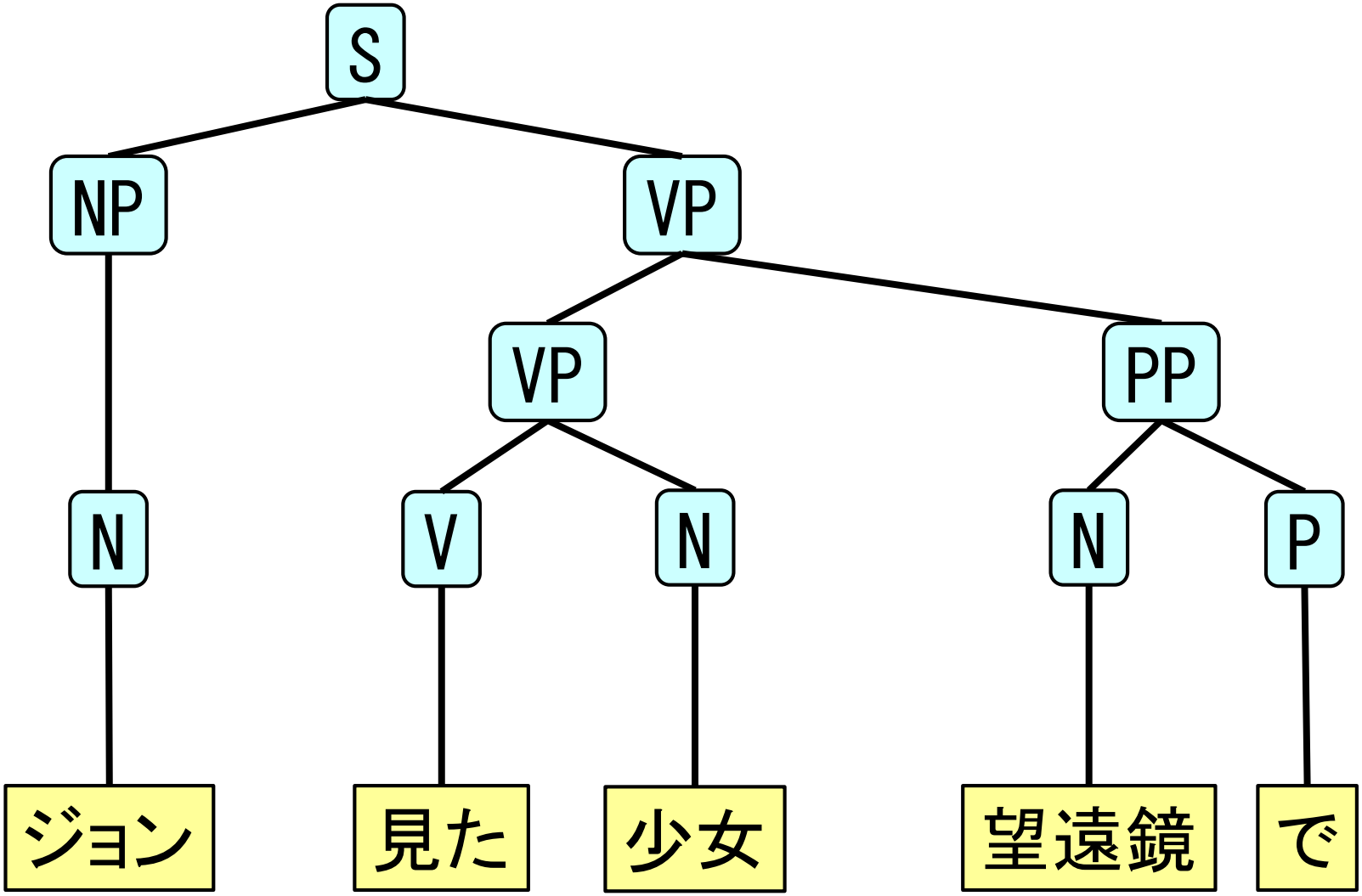
構文木の変換による翻訳



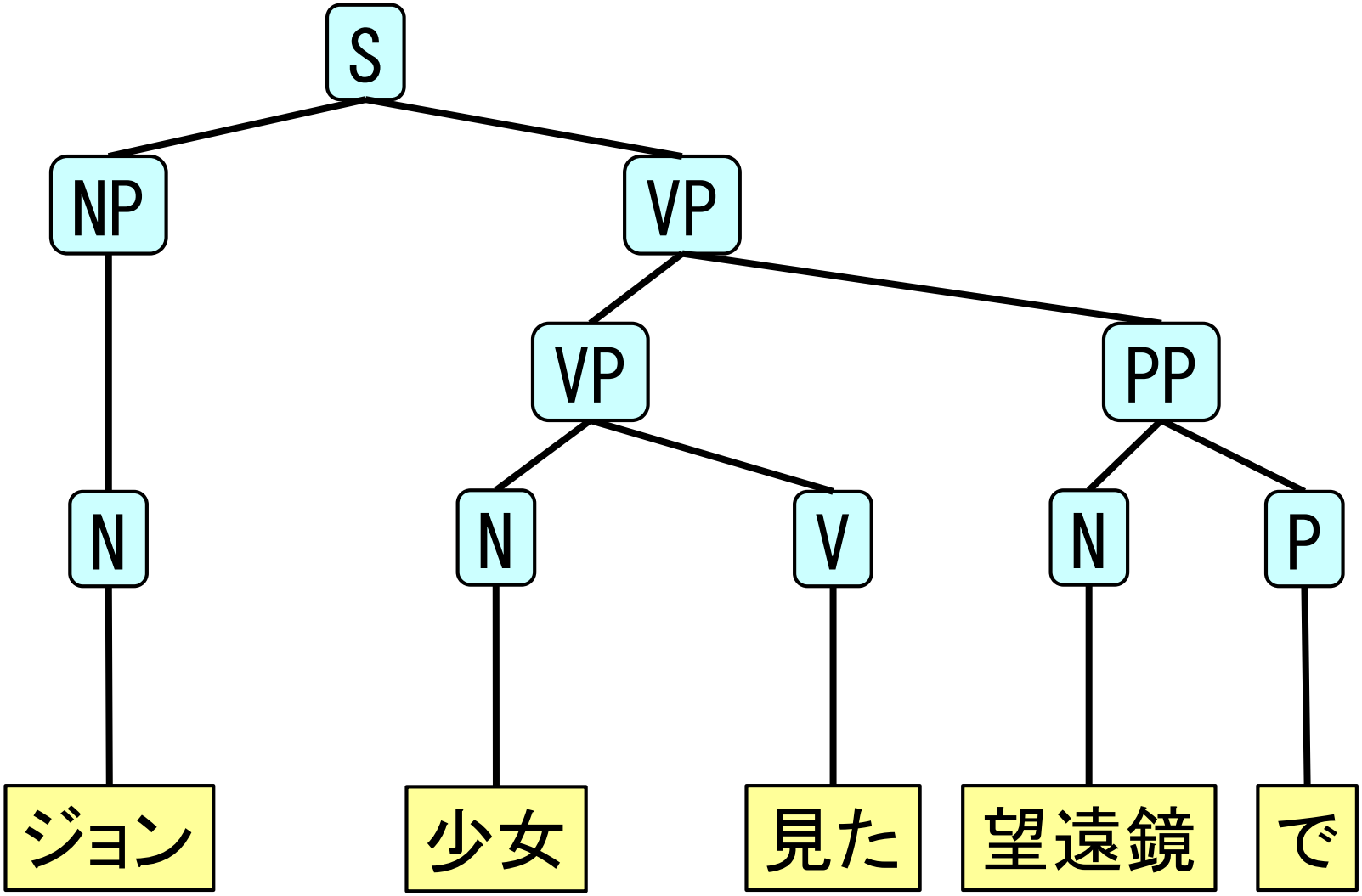
構文木の変換による翻訳



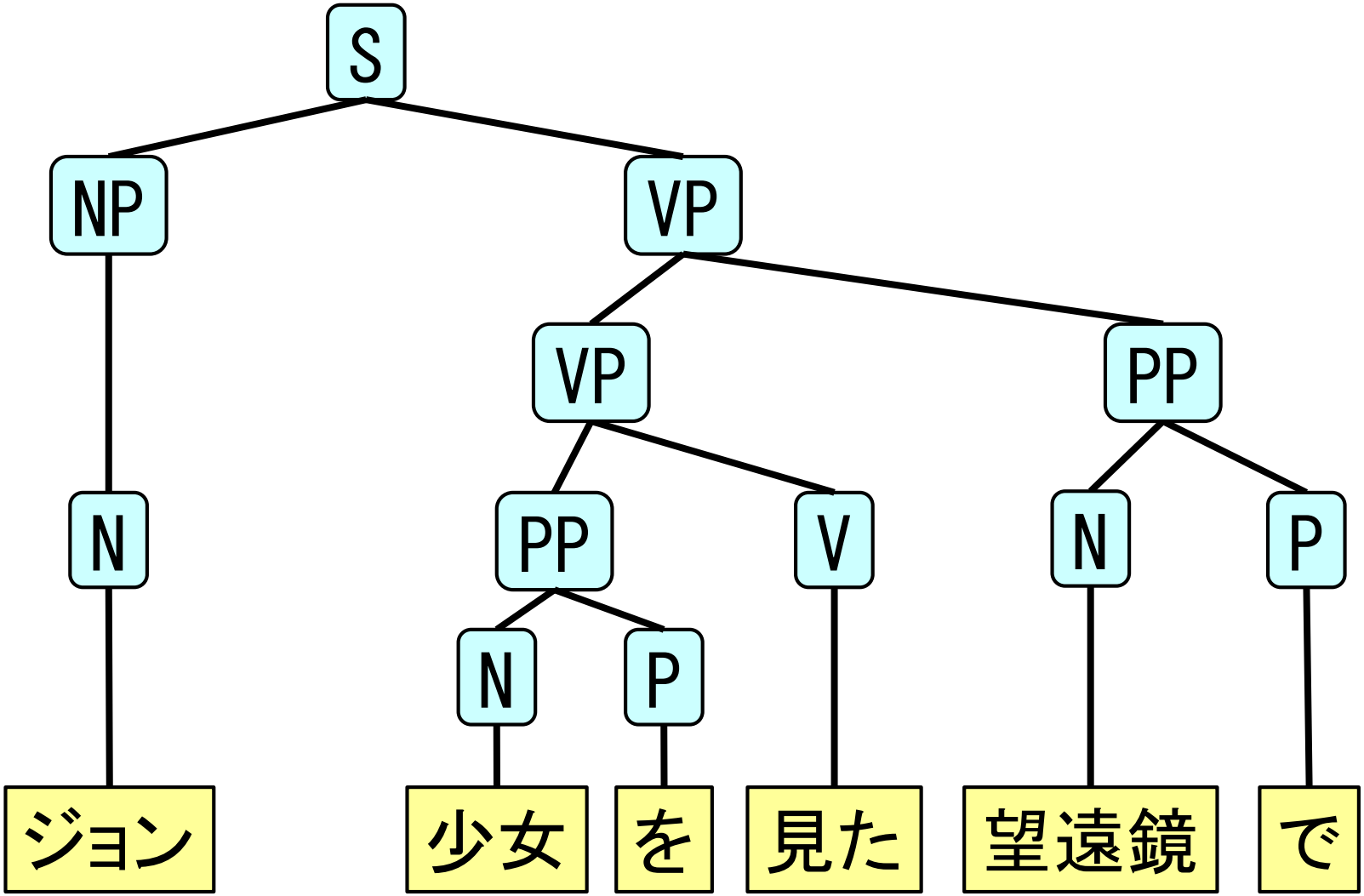
構文木の変換による翻訳



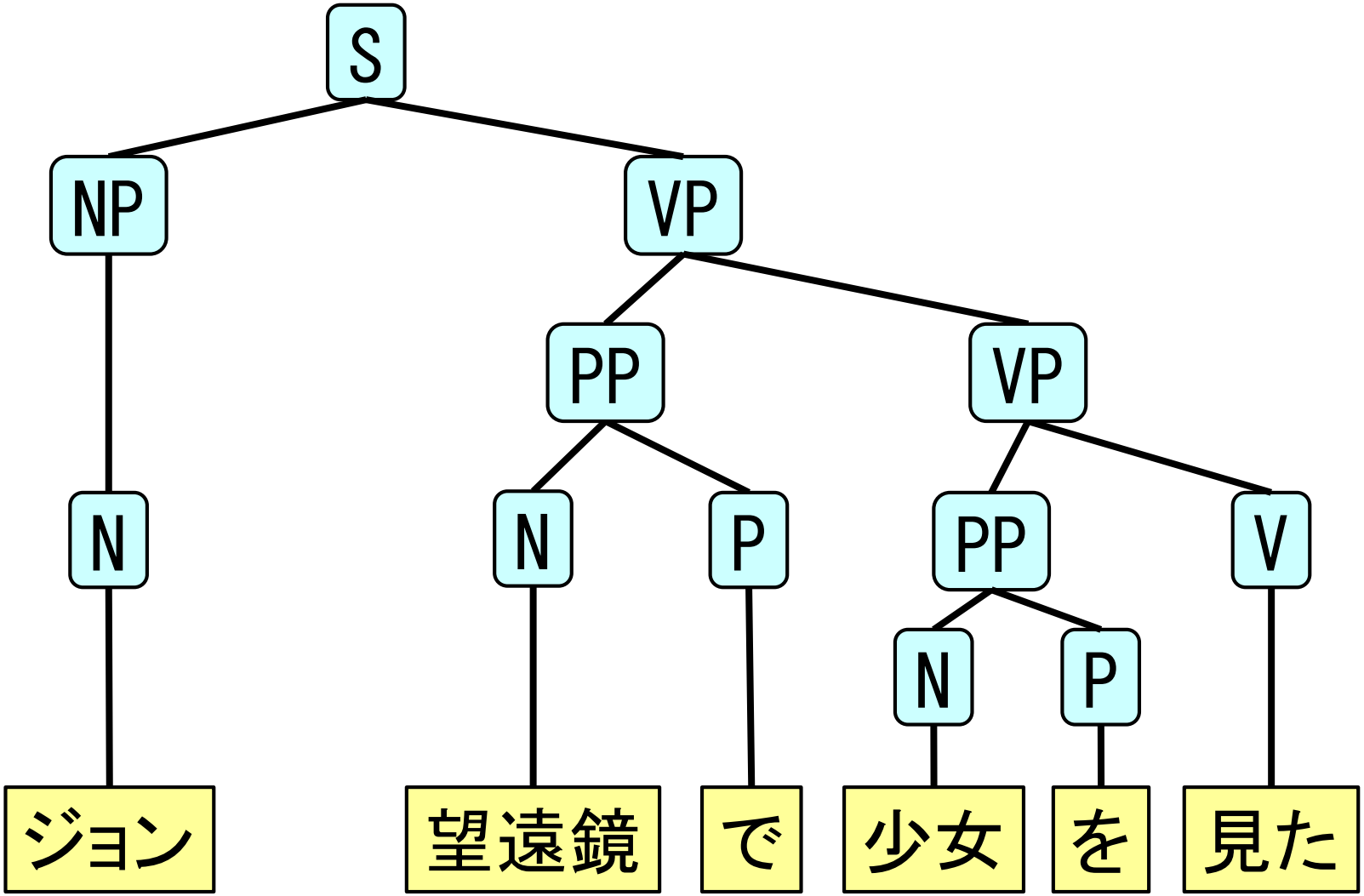
構文木の変換による翻訳



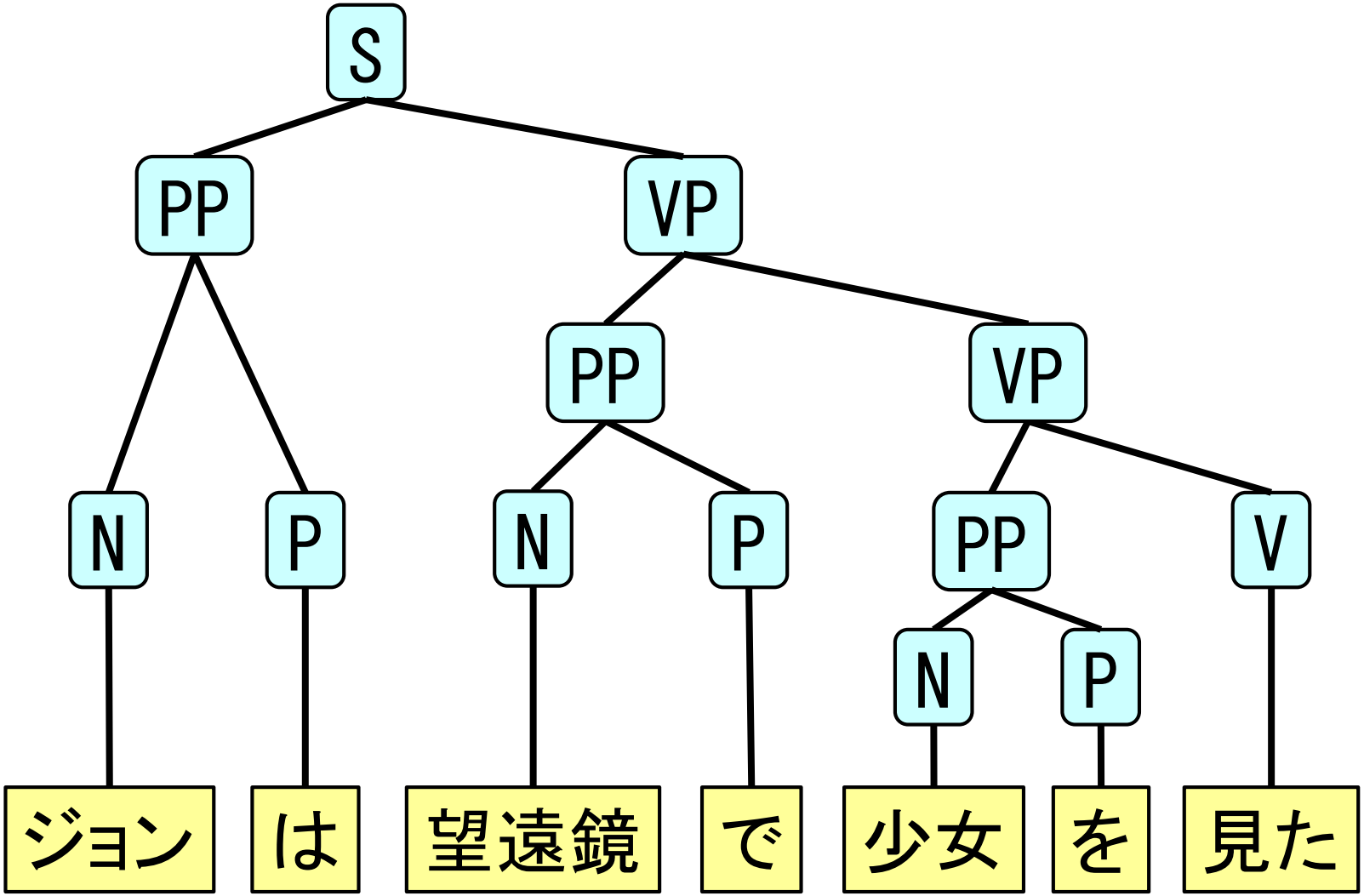
構文木の変換による翻訳



構文木の変換による翻訳



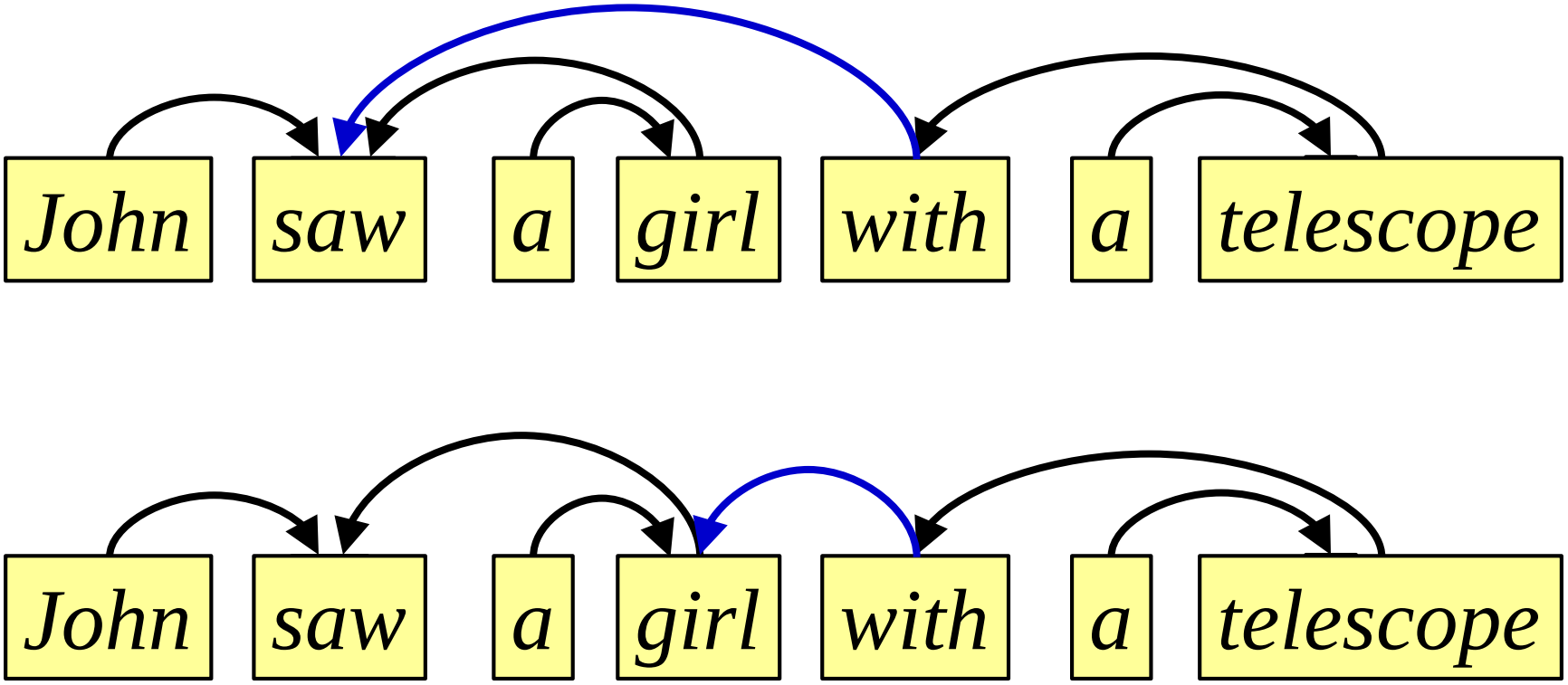
構文木の変換による翻訳



構文解析

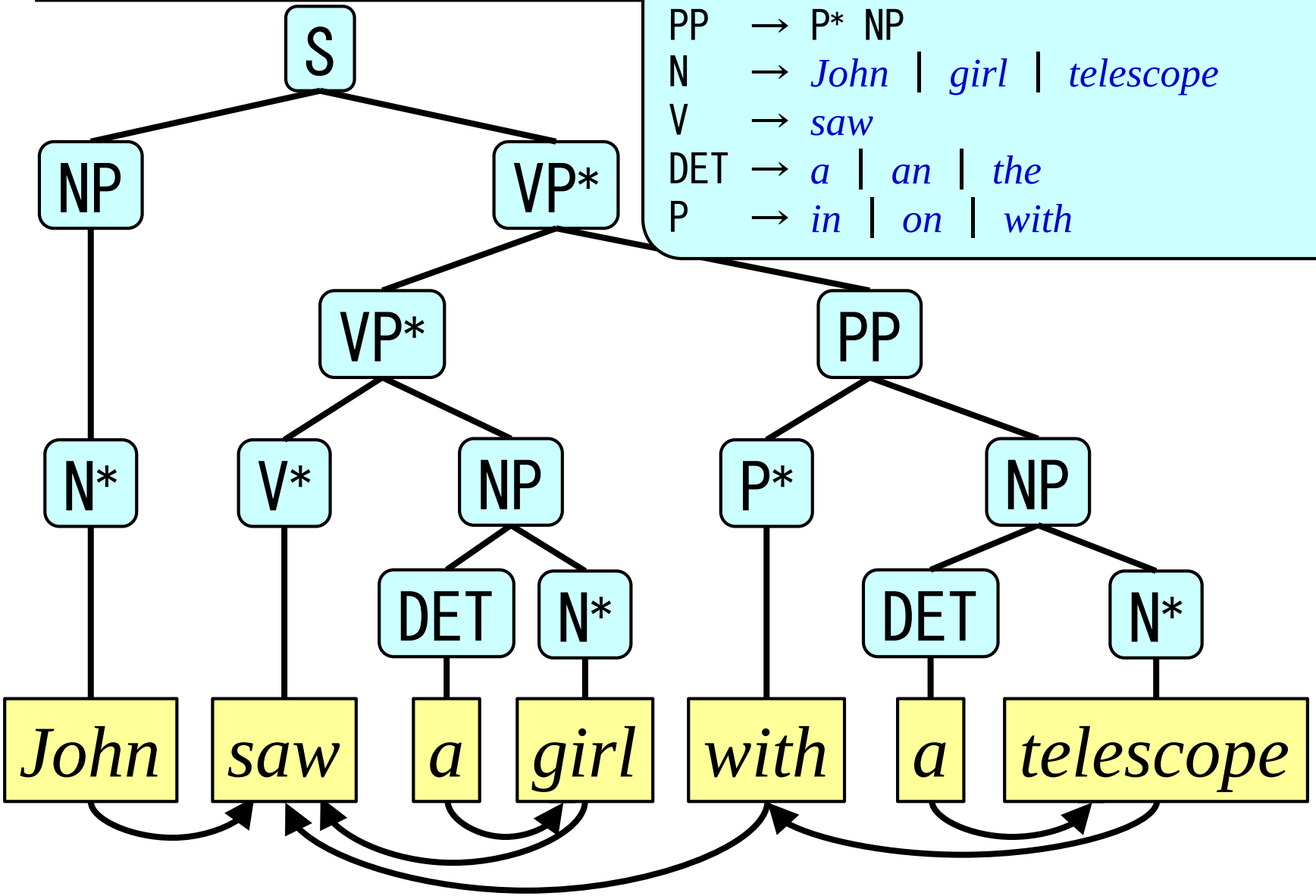
依存文法

依存文法 (Dependency Grammar)



主辞付き構文木

S → NP VP*
NP → N* | DET N* | ADJ N* | NP* PP
VP → V* | V* NP | VP* PP
PP → P* NP
N → *John* | *girl* | *telescope*
V → *saw*
DET → *a* | *an* | *the*
P → *in* | *on* | *with*



文節

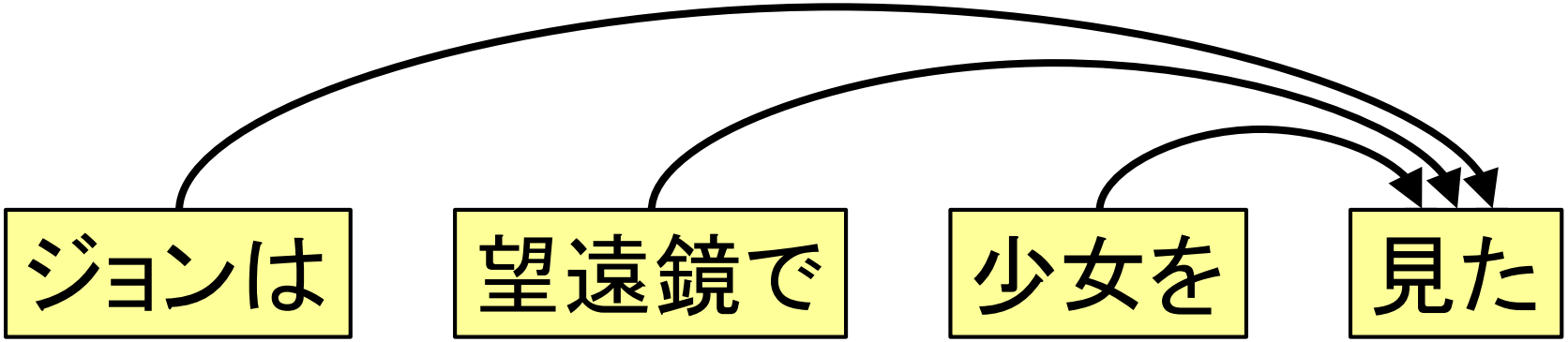
- 日本語において、

  一つの自立語　＋　0個以上の付属語

  からなる単位

ジョンは　望遠鏡で　少女を　見た

日本語の依存関係

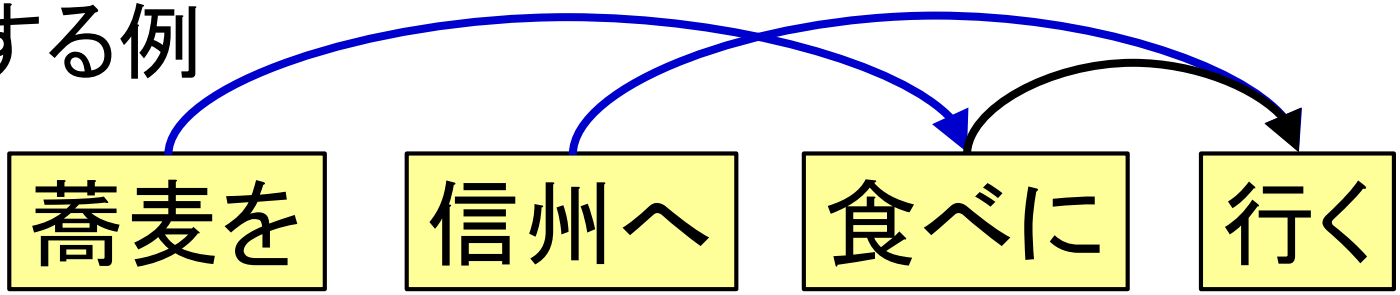


三つの前提条件

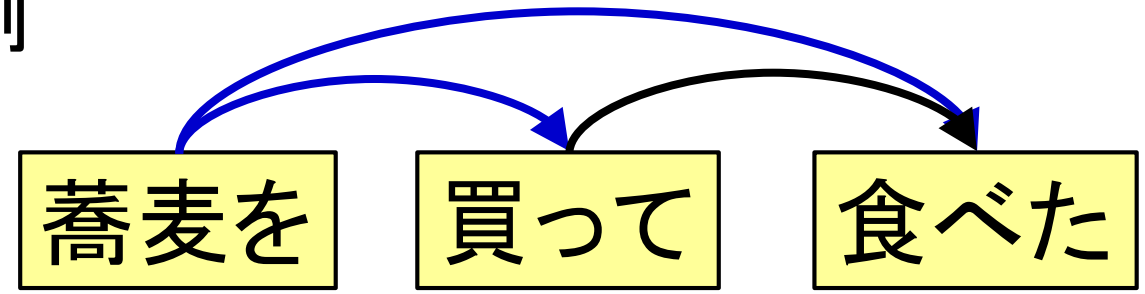
- 非交差性
- 係り先の唯一性
  - ➢ただし、文末の文節のみ係り先がない（ゼロ）
- 後方修飾性

例外

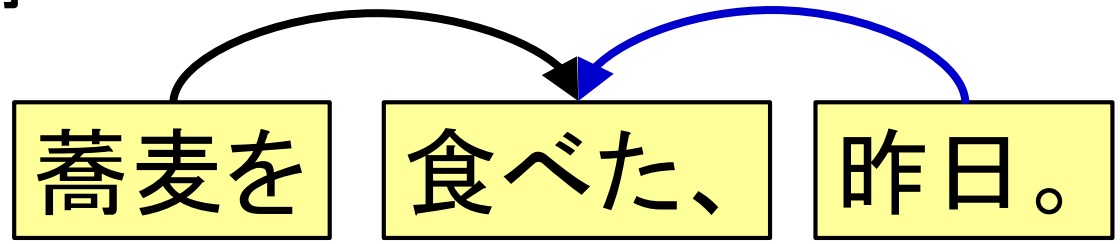
- 交差する例



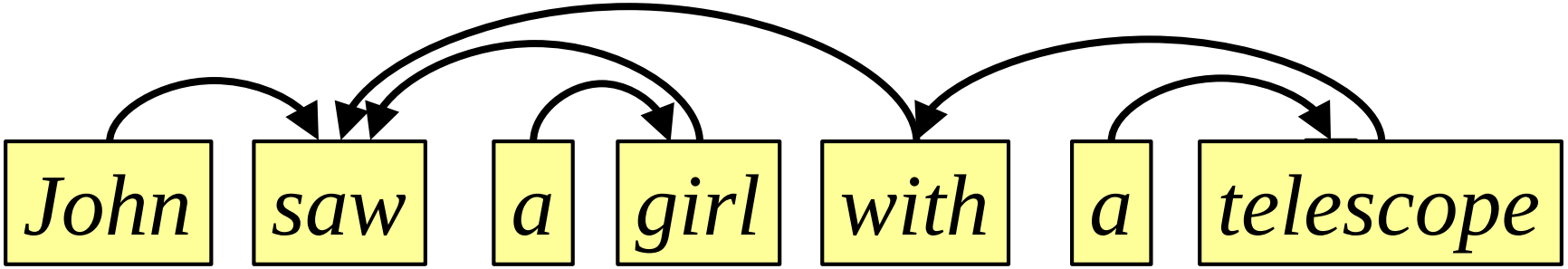
- 複数に係る例

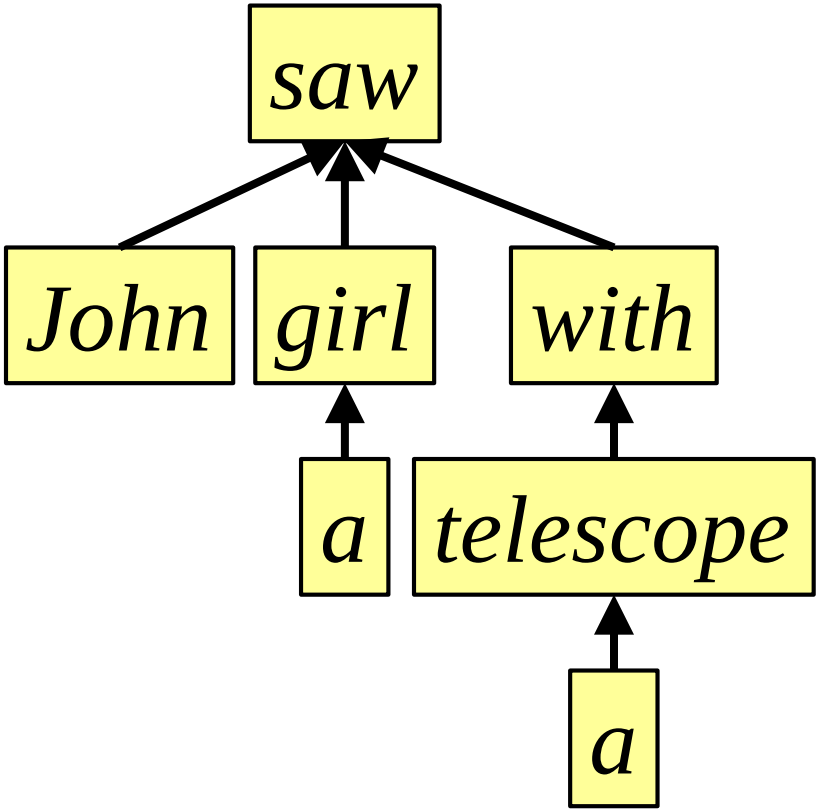


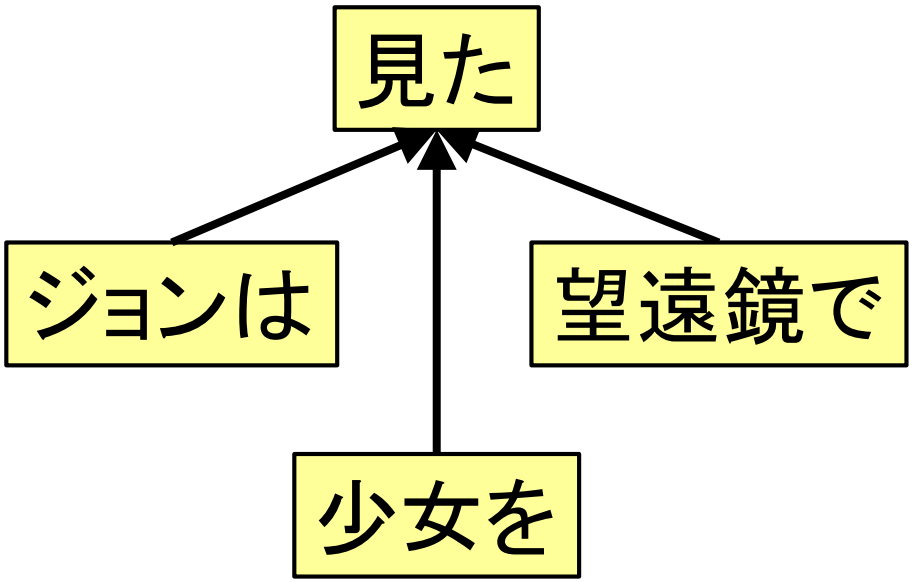
- 前方に係る例



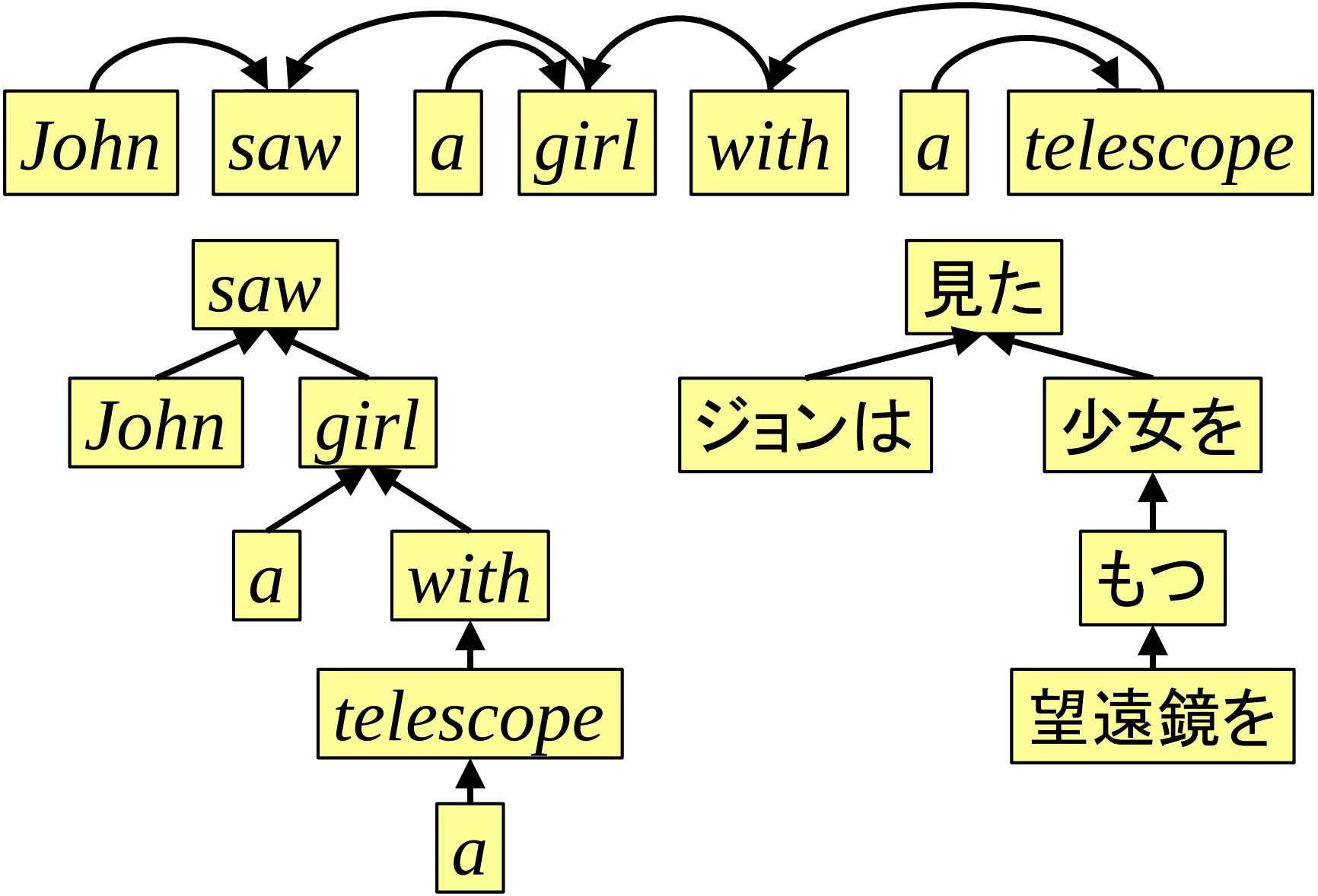
依存文法による翻訳


John
saw
a
girl
with
a
telescope


saw
John
girl
with
a
telescope
a


見た
ジョンは
望遠鏡で
少女を

依存文法による翻訳　別の例



チャンキング (Chunking)

浅い構文解析 (shallow parsing) とも

- 英語
  - 名詞句や動詞句をまとめる
- 日本語
  - 文節に区切る
  - 名詞句や動詞句の抽出

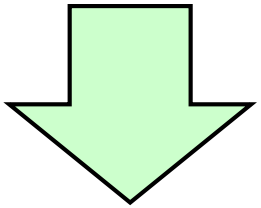
統計的構文解析

- 確率文脈文法
  - 規則に確率を付与
  - 文が生成される確率は、適用した確率の積
    - 生成確率が最大の構文を出力

```
S    → NP VP   (1.0)
NP   → N       (0.2)
     → DET N   (0.3)
     → ADJ N   (0.2)
     → NP PP   (0.3)
```

Treebank

- 構文的な構造が付与されたコーパス
  - Penn Treebank
  - 京都テキストコーパス



コーパスから文法を作成

意味解析

格文法(Case Grammar) [Fillmore, 96]

- 表層格と深層格
- 必須格(obligatory case)と
  任意格(optional case)

**John gave her.*

必須格である与格がない

表層格 (Surface Case)

- 主格(nominative)
- 目的格
  - 対格(accusative)
  - 与格(dative)
- 所有格
  - 属格(genitive)

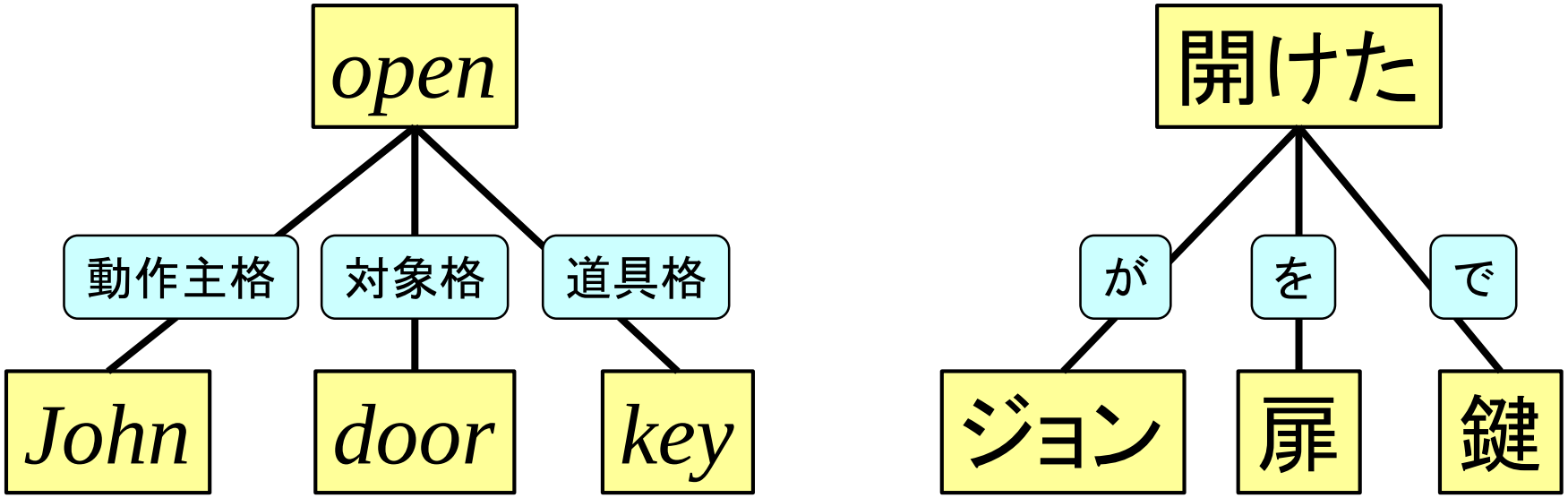
- ガ格
- ヲ格
- ニ格
- デ格
- カラ格
- ヘ格
- ト格
- ヨリ格
- マデ格

深層格 (Deep Case)

- 動作主格(Agent)
- 対象格(Object)
- 目標格(Goal)
- 道具格(Instrument)
- 場所格(Location)
- 時間格(Time)
- 経験者格(Experiencer)
- 源泉格(Source)

格文法による翻訳

John opened the door with the key.



格形態と文法関係のずれ

- 主格でない「が」

  彼女は花が好きだ

- 目的格でない「を」

  公園を歩く

  橋を渡る

表層格から深層格へ

曖昧性がある

対象格	ゲームで遊ぶ	*play a game*
場所格	公園で遊ぶ	*play at the park*
道具格	おもちゃで遊ぶ	*play with a toy*
？	一人で遊ぶ	*play alone*

格フレーム (Case Frame)

- 単語の共起に関する知識
- 動詞の場合、格への制約

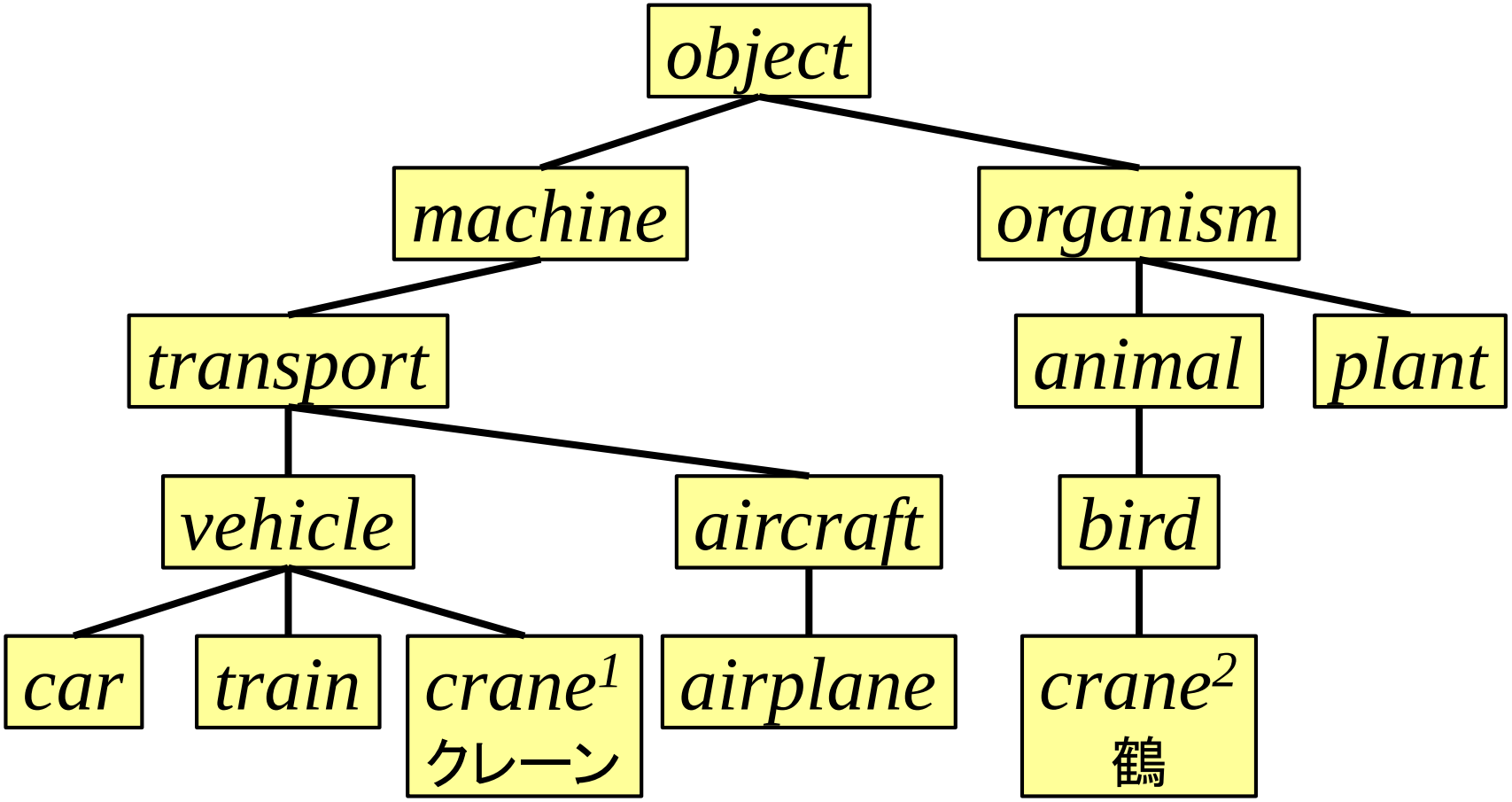
eat:
- 食べる
  (subj, 人間, 動作主)
  (obj, 食物, 対象)

fly:
- 飛ぶ
  (subj, {鳥,航空機}, 動作主)

take:
- 撮る
  (subj, 人間, 動作主)
  (obj, 写真, 対象)
- 乗る
  (subj, 人間, 動作主)
  (obj, 乗り物, 対象)
- 飲む
  (subj, 人間, 動作主)
  (obj, 薬, 対象)

シソーラス (Thesaurus)

- 上位・下位関係、同義関係などによって単語を分類し体系化したもの



語義曖昧性解消 (Word Sense Disambiguation)

格フレームとシソーラスを利用

A crane flies.

fly:
- 飛ぶ
  (subj, {鳥,航空機}, 動作主)

He took a bus.

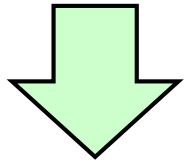
take:
- 撮る
  (subj, 人間, 動作主)
  (obj, 写真, 対象)
- 乗る
  (subj, 人間, 動作主)
  (obj, 乗り物, 対象)
- 飲む
  (subj, 人間, 動作主)
  (obj, 薬, 対象)

オマケ

翻字 (transliteration)

- 文字から文字への変換
  - 音から文字への場合は転写/音訳（transcript）
- 固有名詞の翻訳に必要

Audrey Hepburn



オードリー・ヘップバーン

翻字の曖昧性

- 同じ文字・同じ発音でも異なる

Canberra → キャンベラ

California → カリフォルニア

- 正書法が定まっていない

spaghetti →

スパゲッティー	スパゲティー
スパゲッティ	スパゲティ
スパゲッテー	スパゲテー
スパゲッテ	スパゲテ

文字の選択

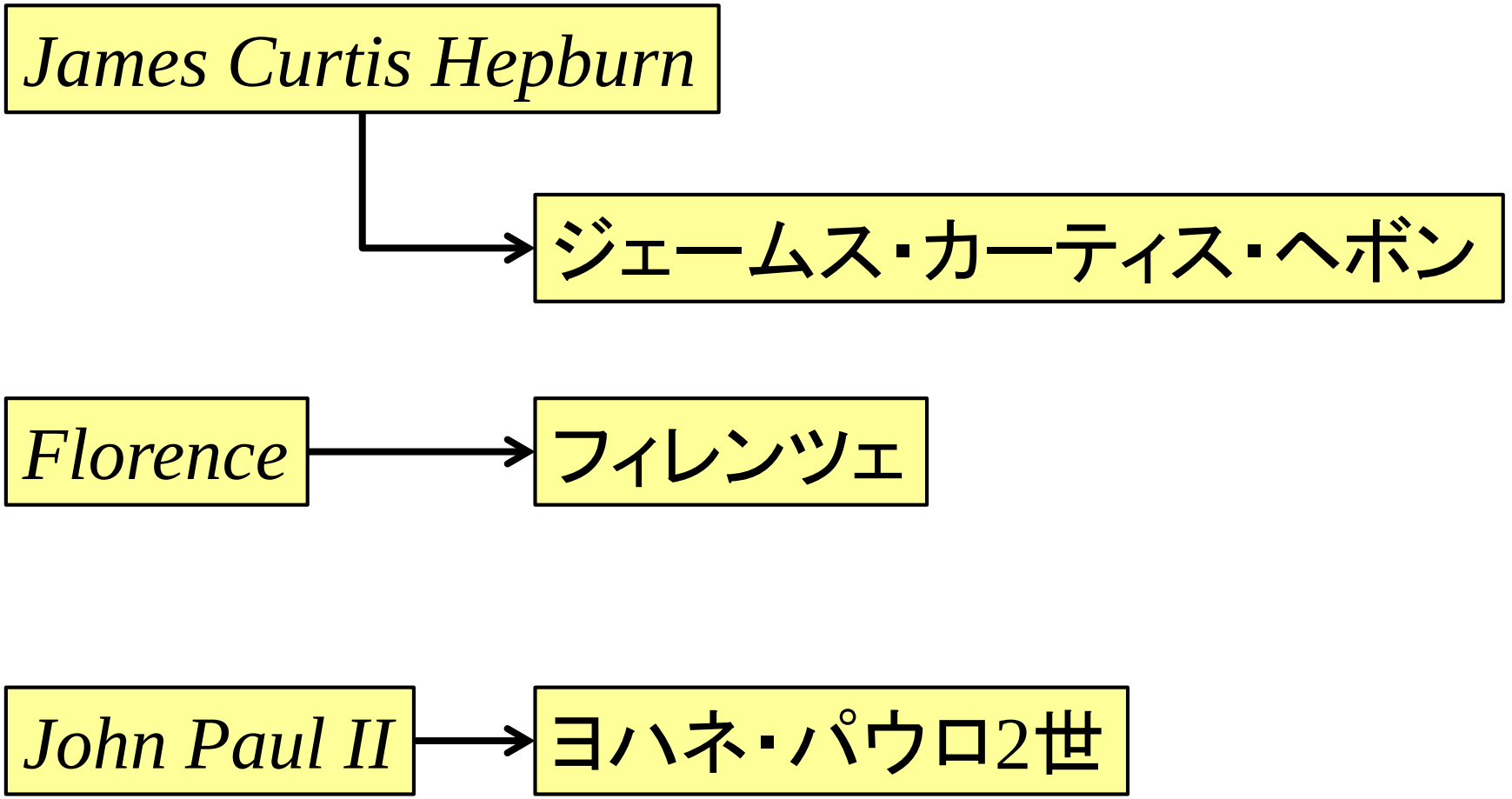
- 中国語への翻字

Coca-Cola → コカコーラ

可口可乐

可口可樂　飲んで楽しい

歴史的・文化的な理由


James Curtis Hepburn
ジェームス・カーティス・ヘボン
Florence
フィレンツェ
John Paul II
ヨハネ・パウロ2世

名前の転写

- John, Jan, Giovanni, Ivan, Johan, Johannes, Ioannes, Hans
- George, Georges, Giorgio, Georg, Georgios, 譲治
- Naomi, Noemi, 奈緒美

(谷崎潤一郎 『痴人の愛』)

ヘボン式ローマ字表記

- マッチ　matchi
- 新聞　shimbun
- 新庄　Shinjō　または SHINJOO
- 譲治　Jōji　または JOOJI

ローマ字表記での長音の扱い

- 「えー」 ex.映画

  eiga
- 「おー」 ex. 伊藤（いとう）、大野（おおの）
  - 訓令式　　Itô, Ôno
  - ヘボン式　　Itō, Ōno
  - 駅名　　Itō, Ōno
  - パスポート　　Ito, Ono, 特例 Itoh/Itou, Ohno/Oono
  - 道路標識　　Ito, Ono

読み仮名と一致しない例:　　講師　kōshi
子牛　koushi