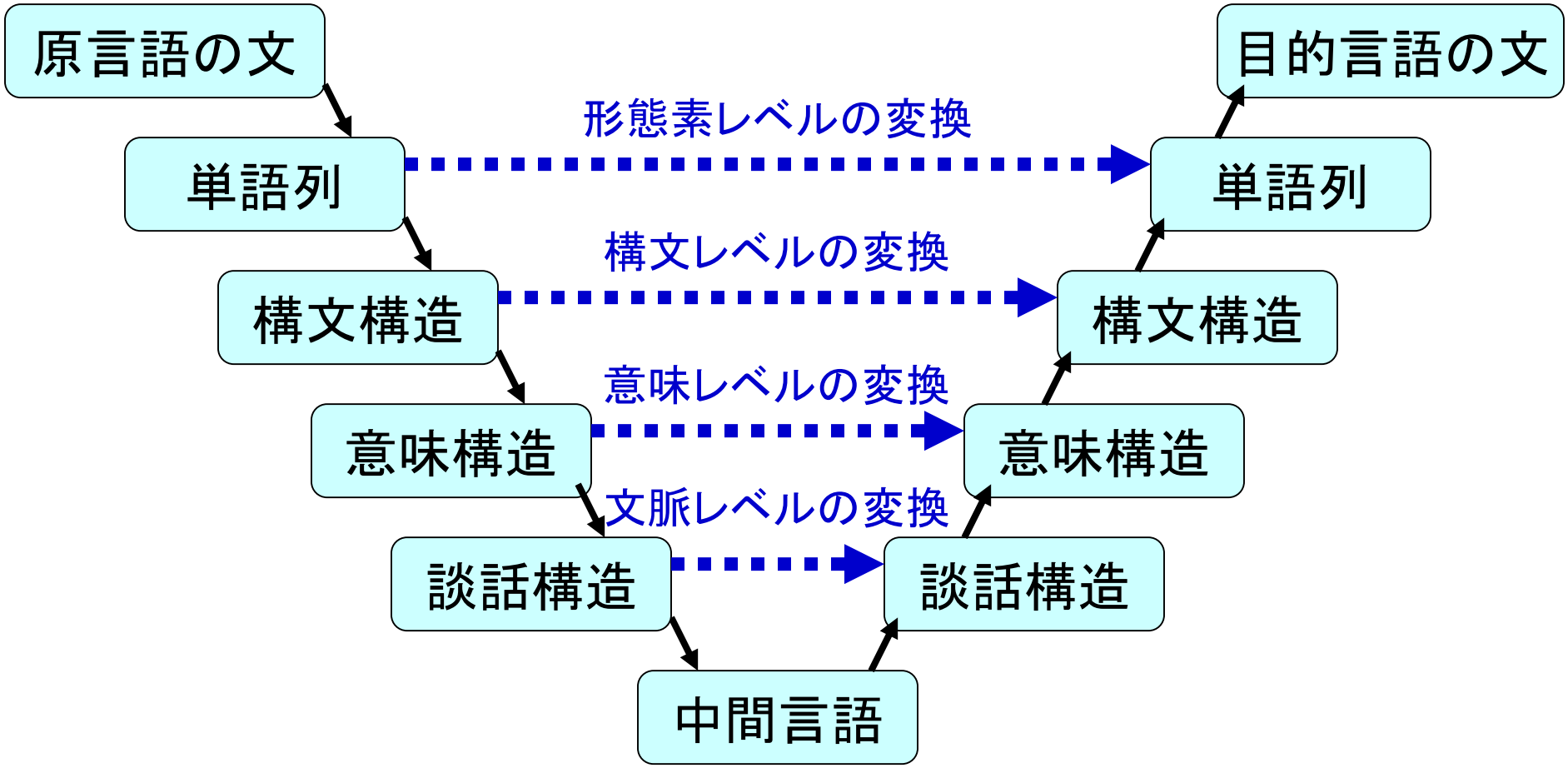
共生社会特論

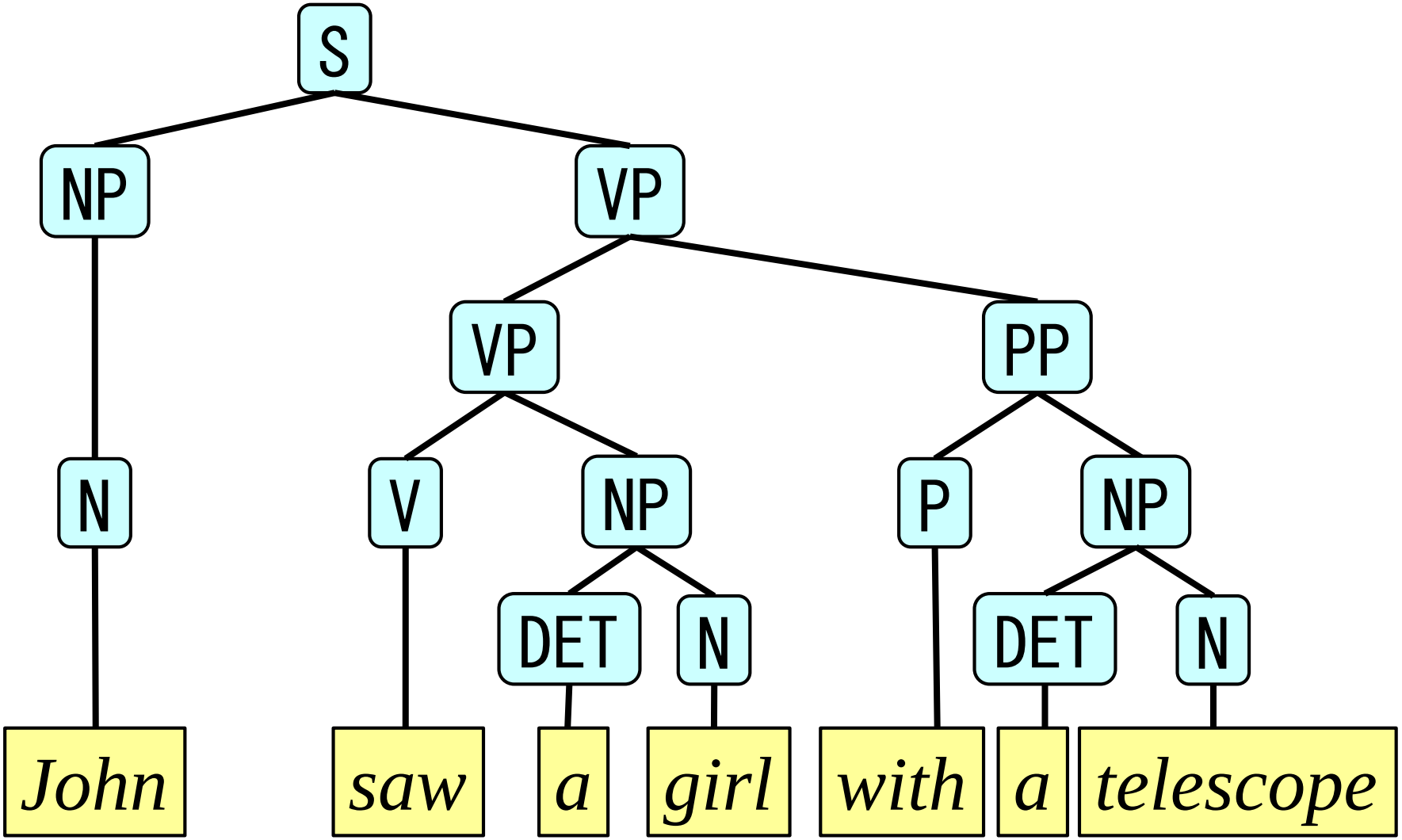
第3回　統計的機械翻訳

2016年12月13日

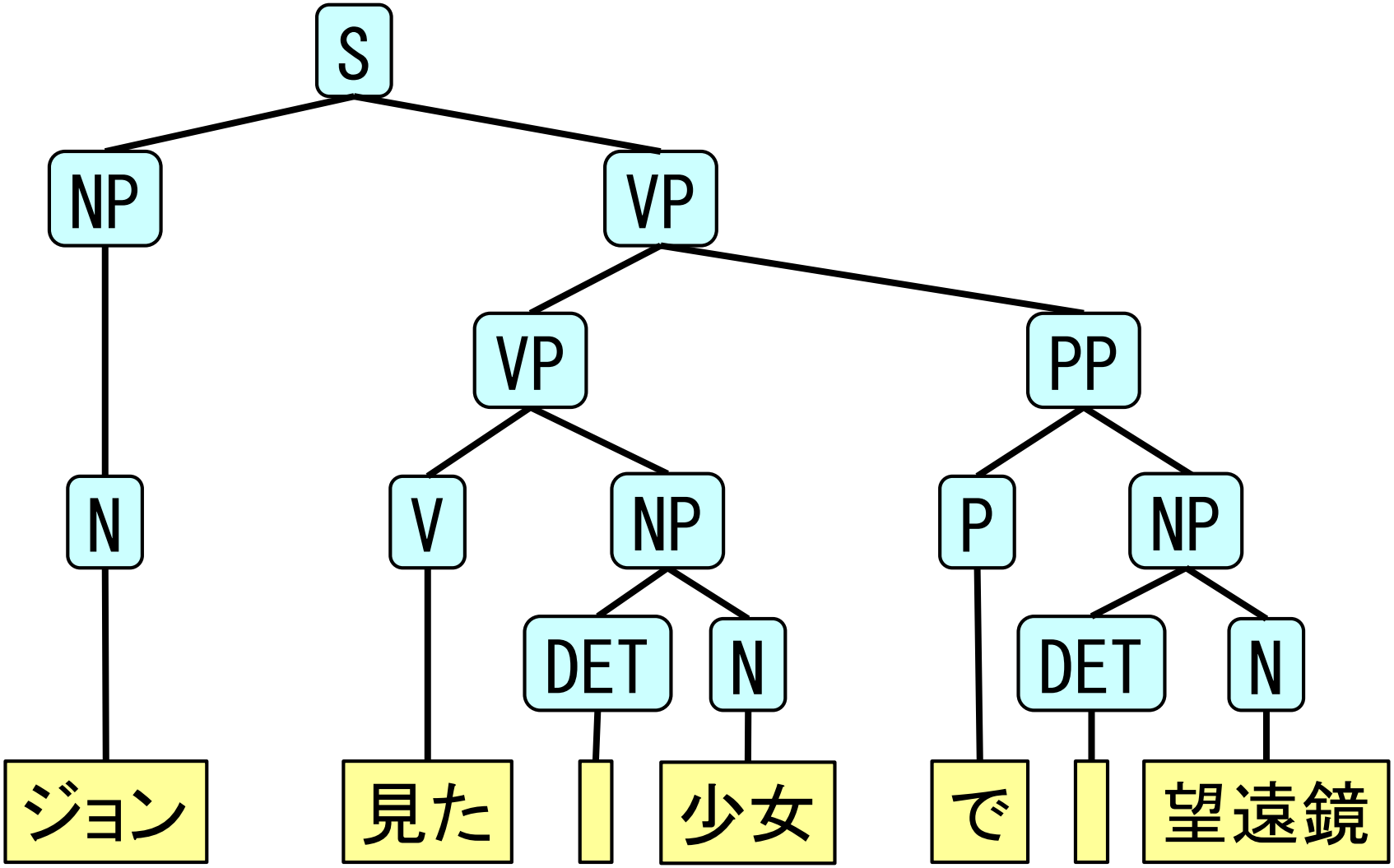
機械翻訳における処理レベル



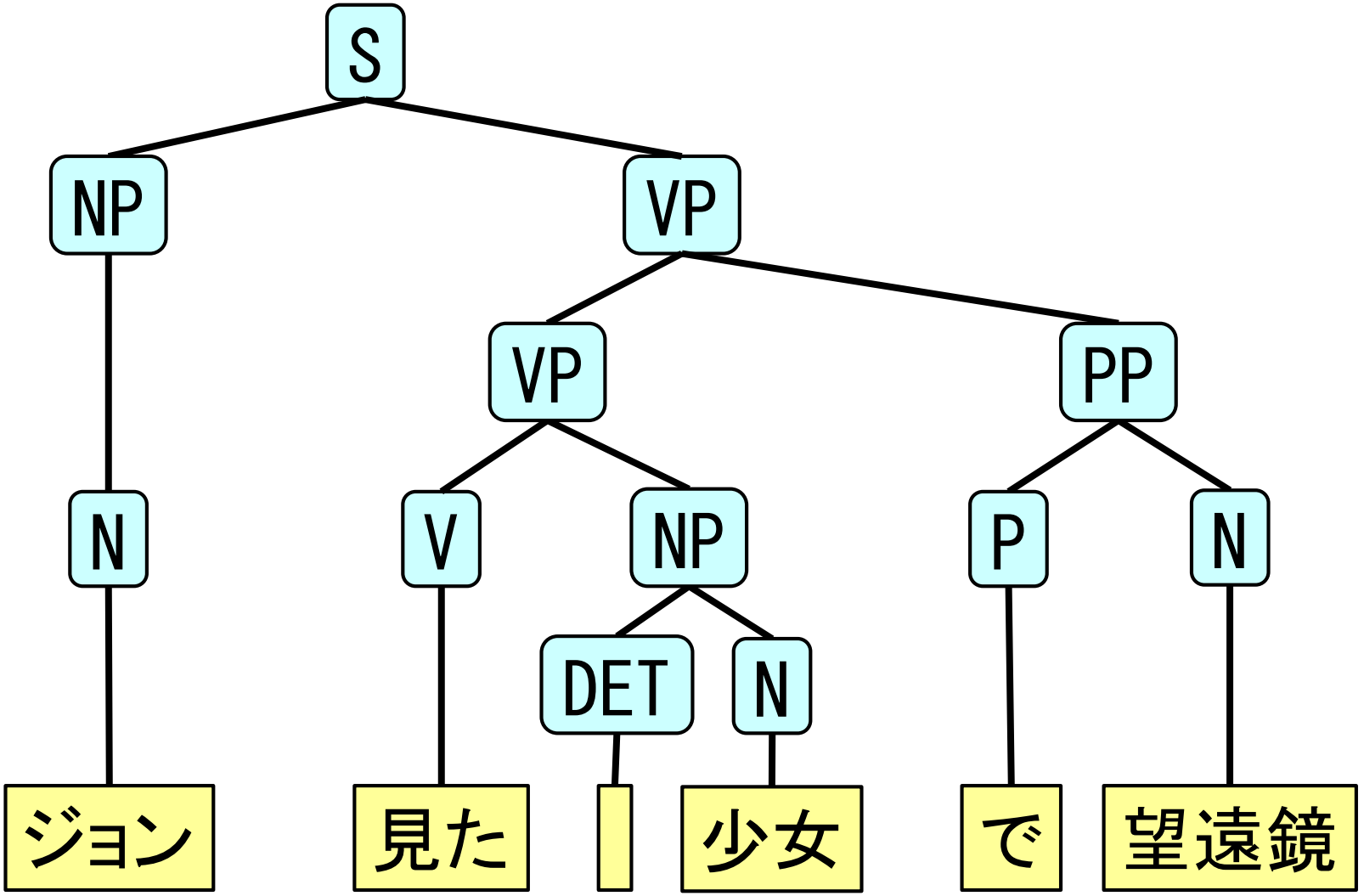
ルールベース翻訳


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

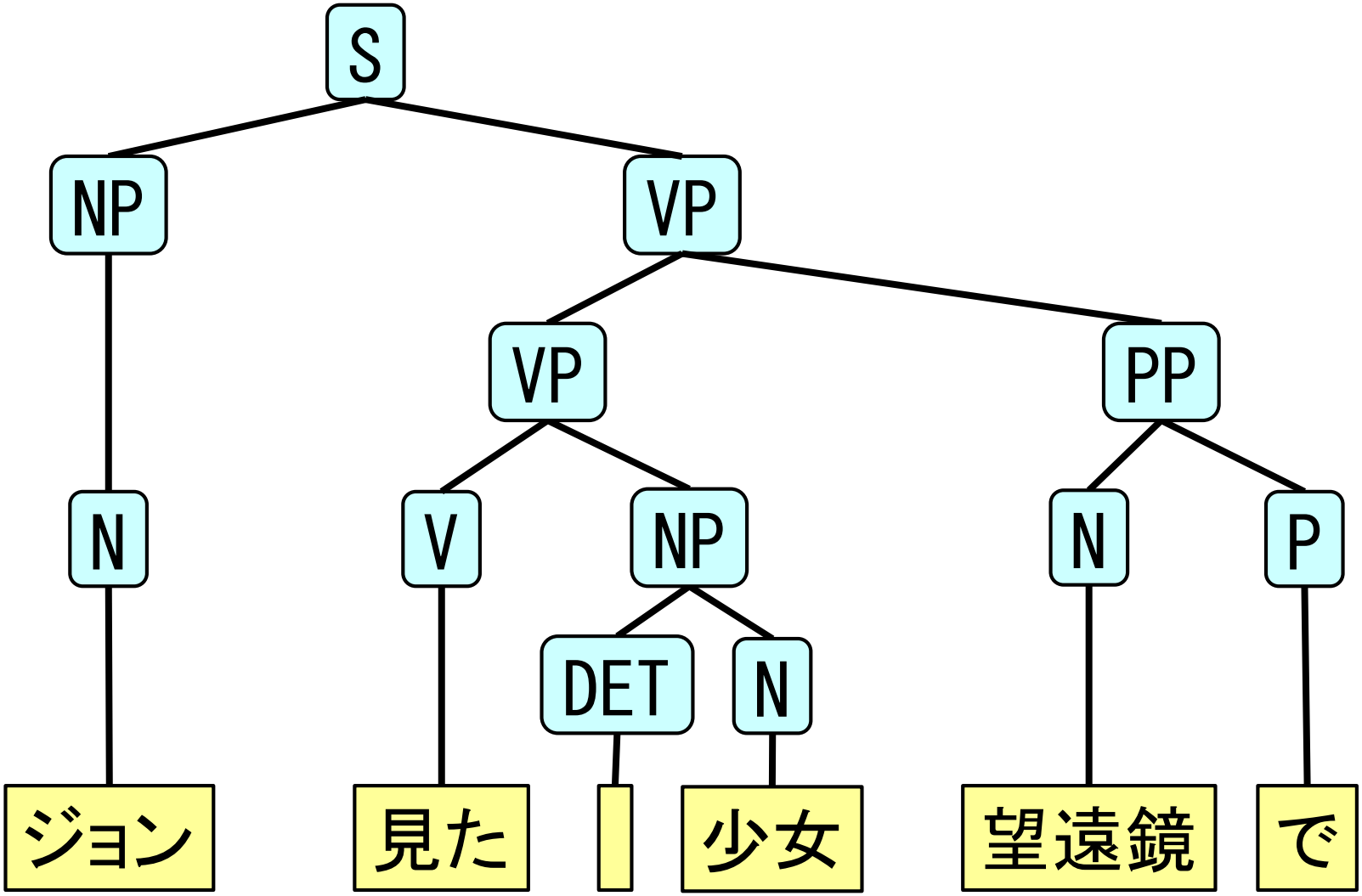
構文木の変換による翻訳



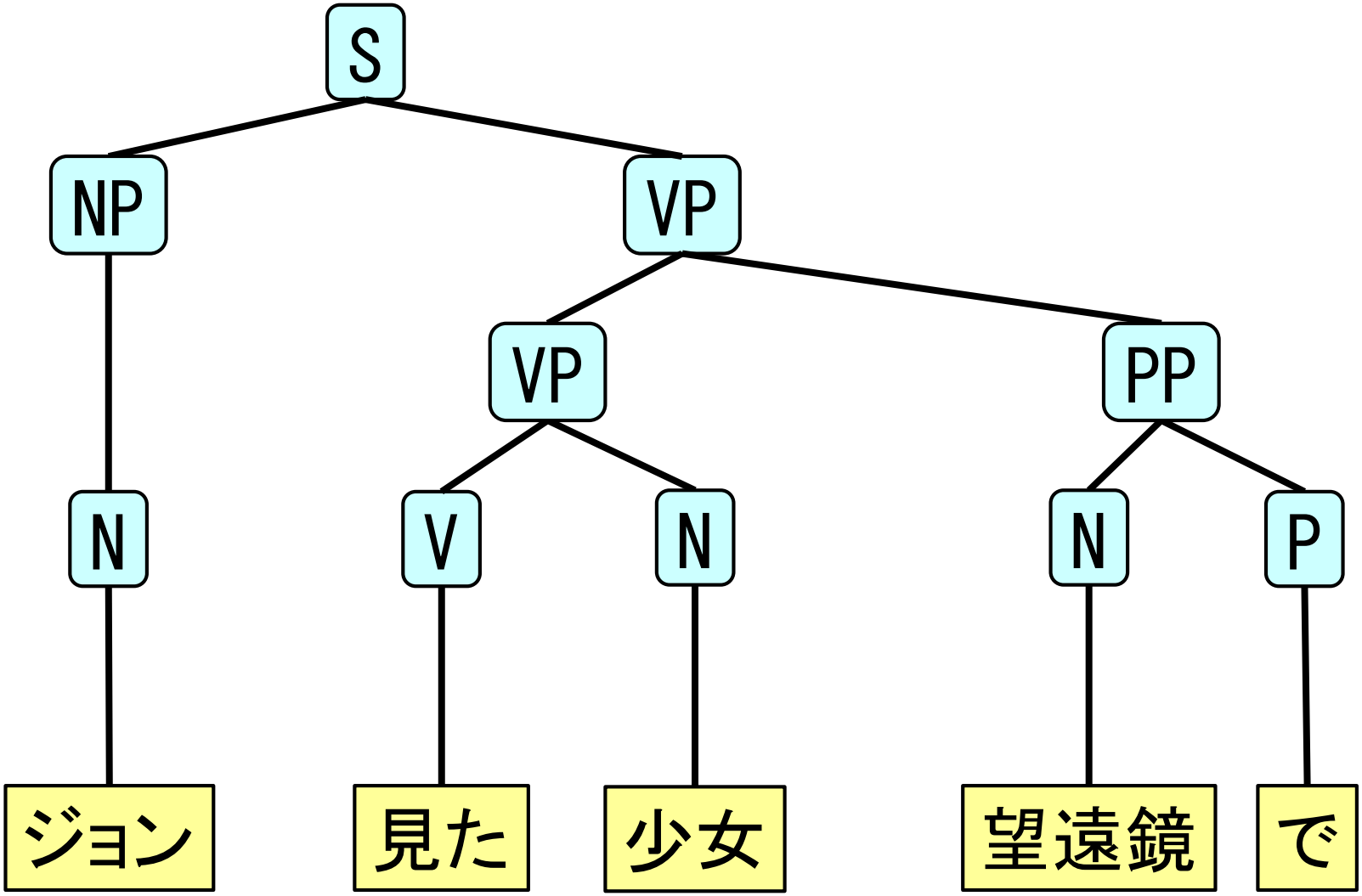
構文木の変換による翻訳



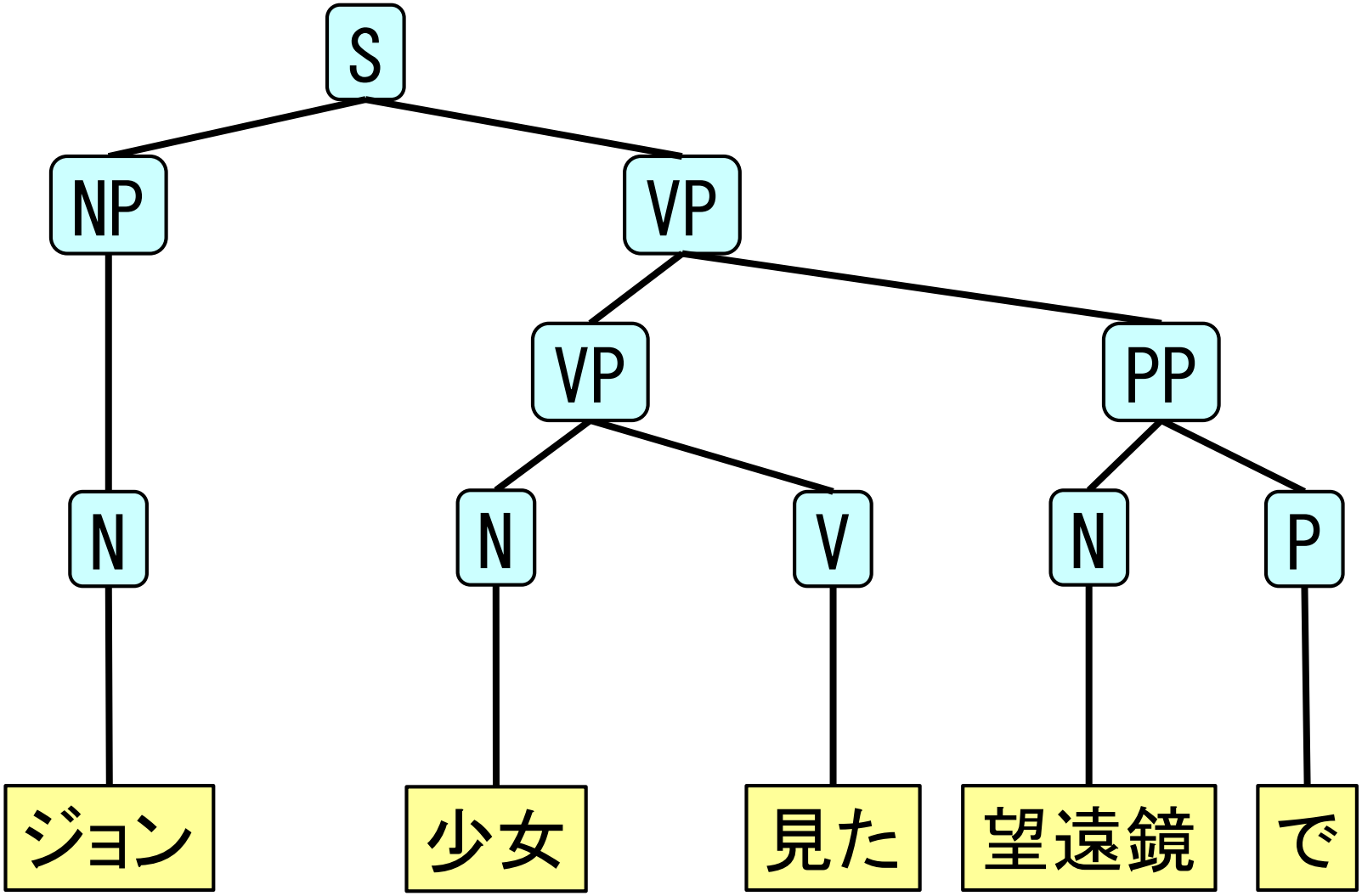
構文木の変換による翻訳



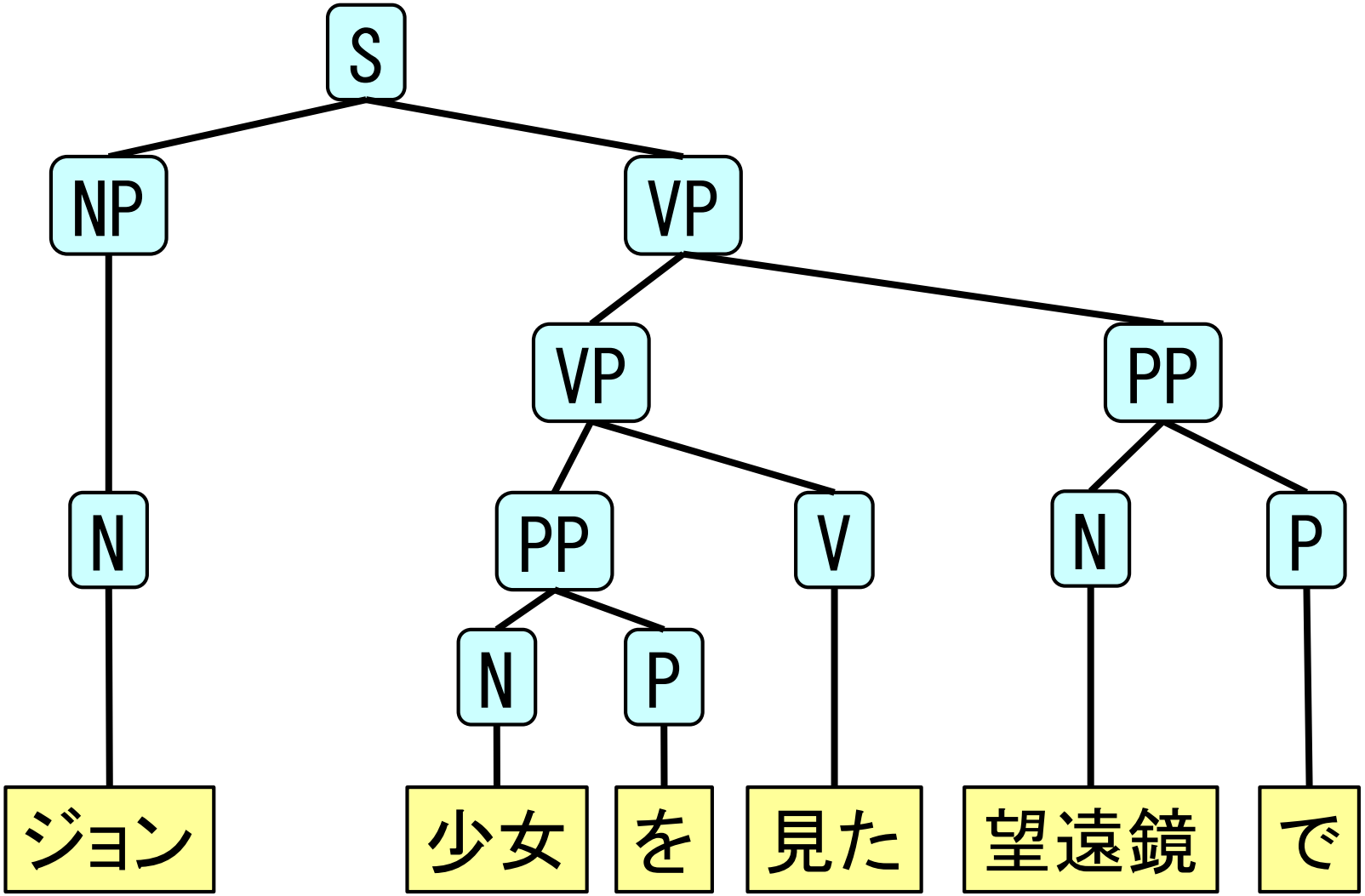
構文木の変換による翻訳



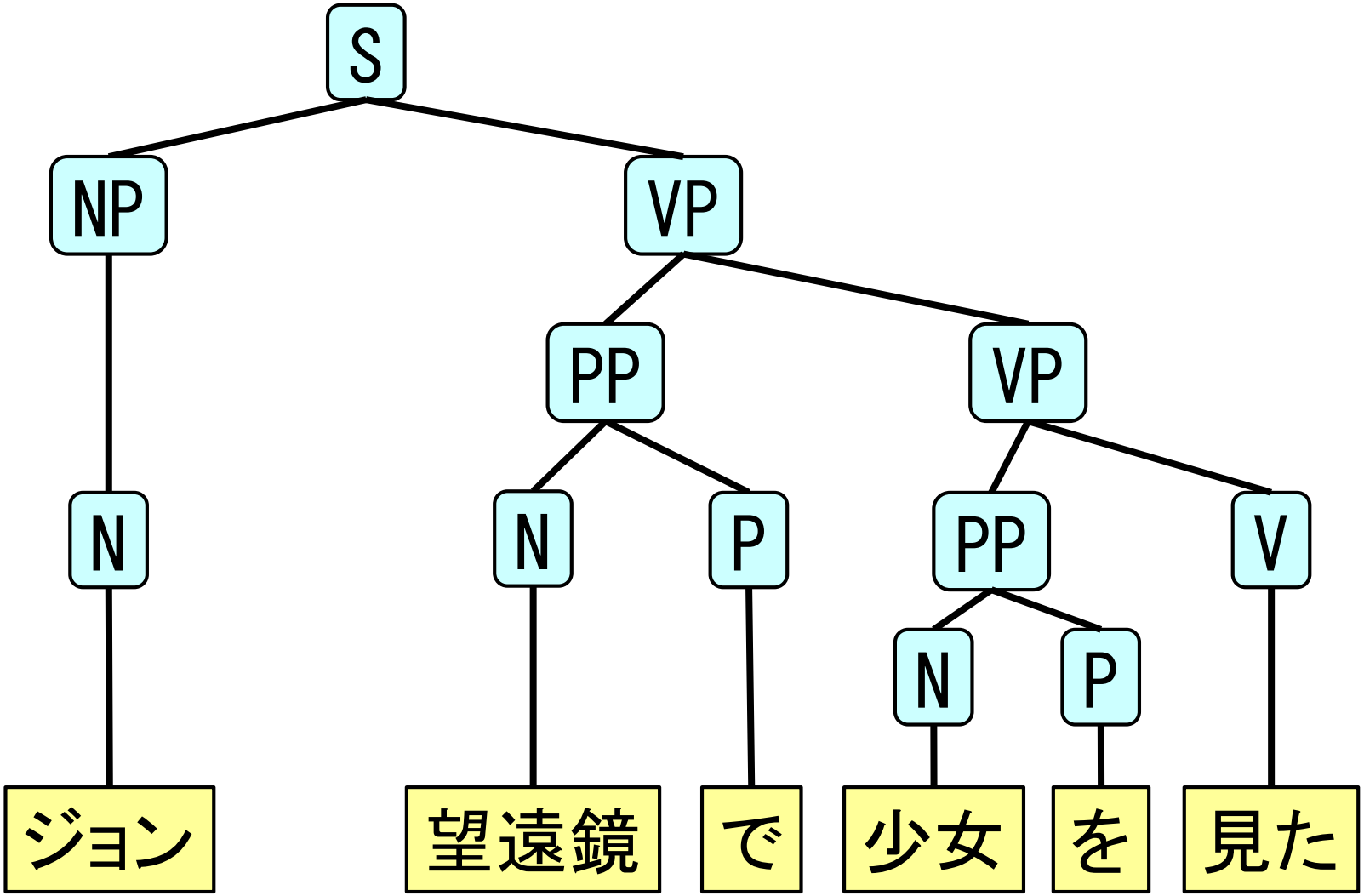
構文木の変換による翻訳



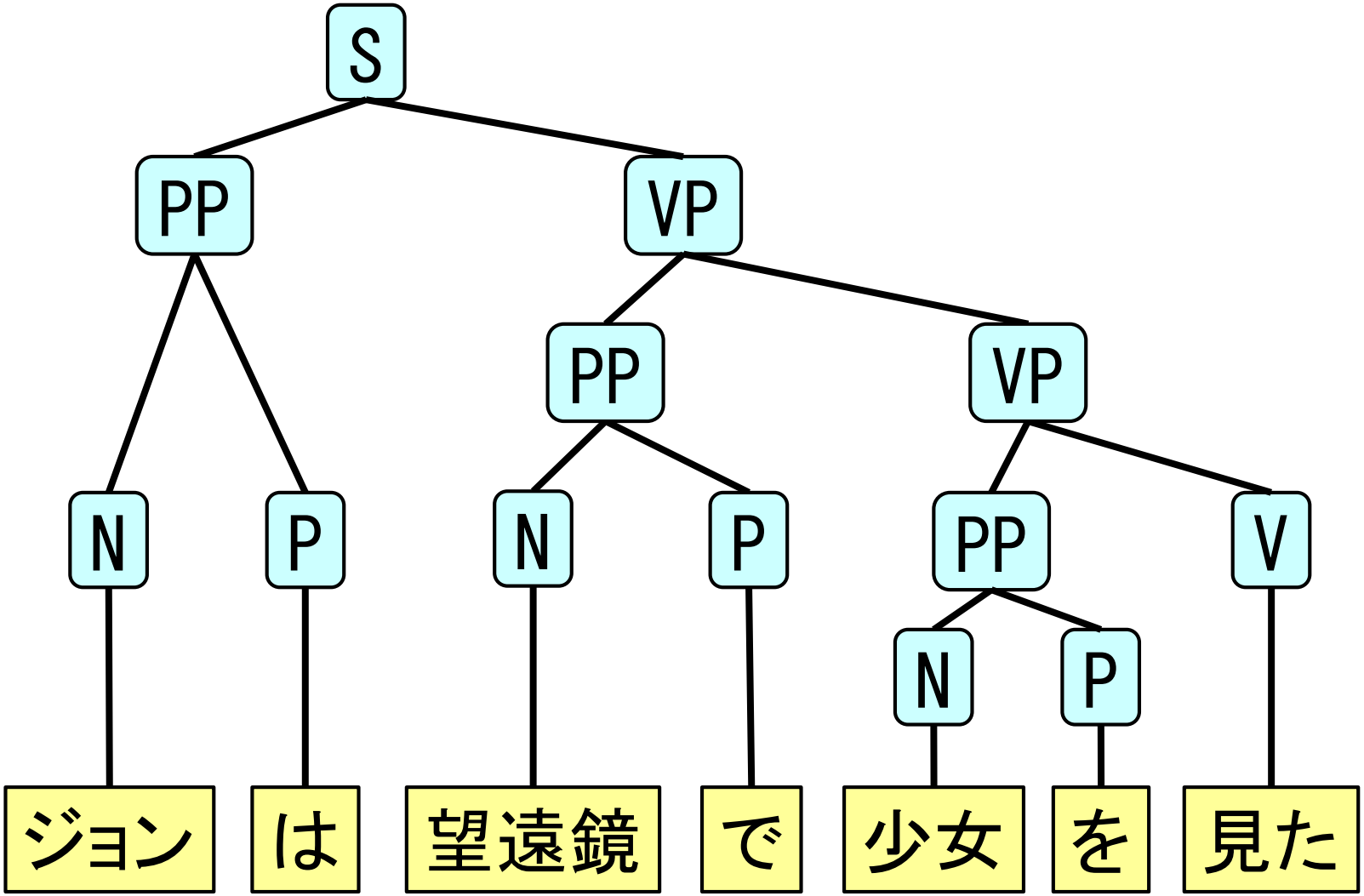
構文木の変換による翻訳



構文木の変換による翻訳



構文木の変換による翻訳



ルールベース翻訳の短所

- ルール作成のコストが高い
  - 両言語に関する専門知識
  - 膨大な数のルール（例外処理）
  - 言語ペアごとにルールが必要

統計的機械翻訳

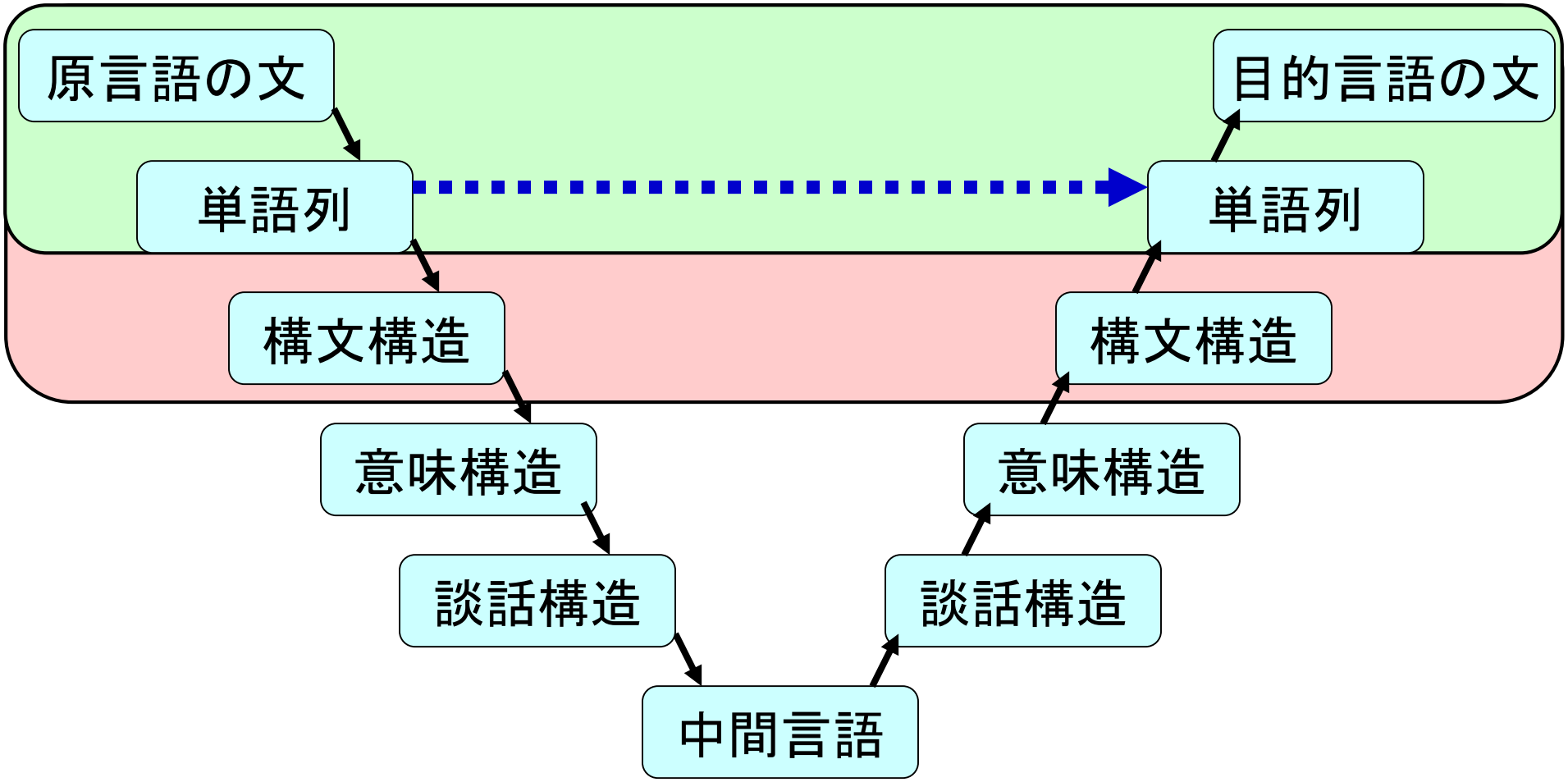
- ルールの自動作成
  - コーパスからの学習
  - 言語に依存しない

対訳コーパス (Bilingual Corpus)

- Parallel Corpus
  - ある文書とそれを翻訳した文書のペア
  - 文と文の対応がついている

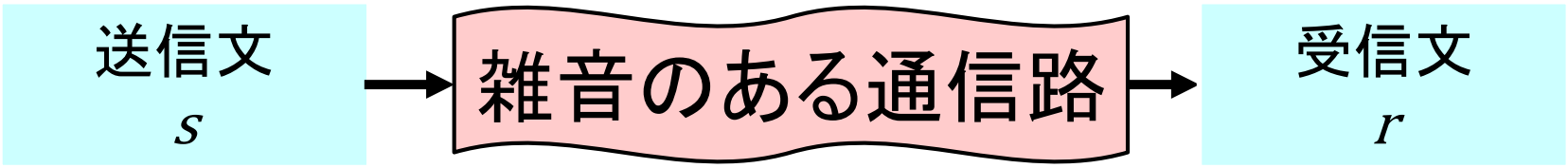
- Comparable Corpus
  - 同じ対象を扱った別言語の文書
    - 新聞記事
    - Wikipedia

統計的機械翻訳の処理レベル

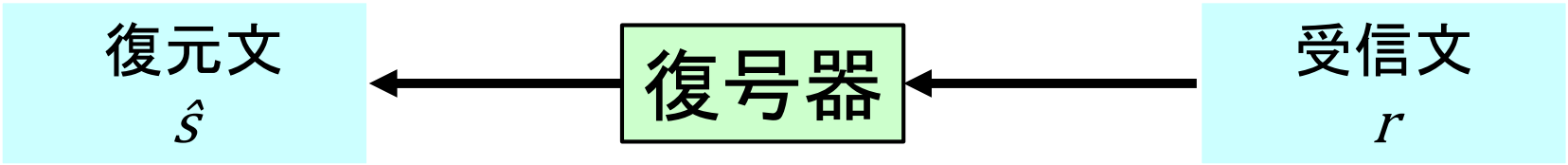


雑音のある通信路モデル (Noisy Channel Model)

- 送信文 s が雑音により r となって届く



- 受信文 r から元の送信文を推測

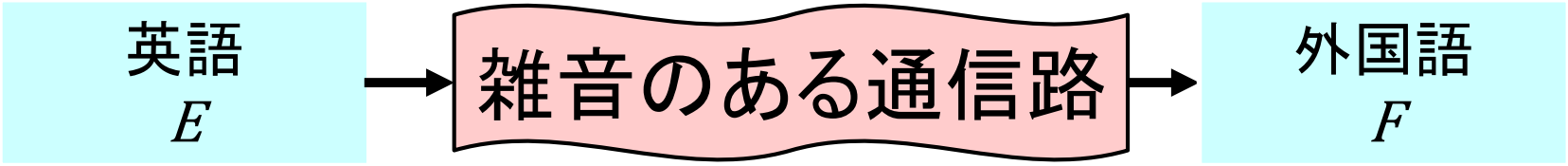


$$\hat{s} = \underset{s}{\operatorname{argmax}} P(s|r)$$

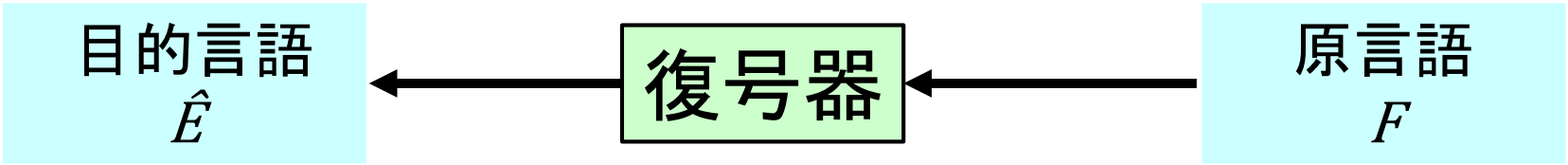
統計的機械翻訳への適用

外国語文から英語文への翻訳

- 英語文 E が雑音により外国語文 F となる



- 復号により英語文 $\hat{E}$ を推測



$$\hat{E} = \underset{E}{\operatorname{argmax}}\, P(E|F)$$

対訳辞書の自動構築

- Word Alignment（単語対応付け）
  - パラレルコーパスから原言語と目的言語の間の単語対応をつける

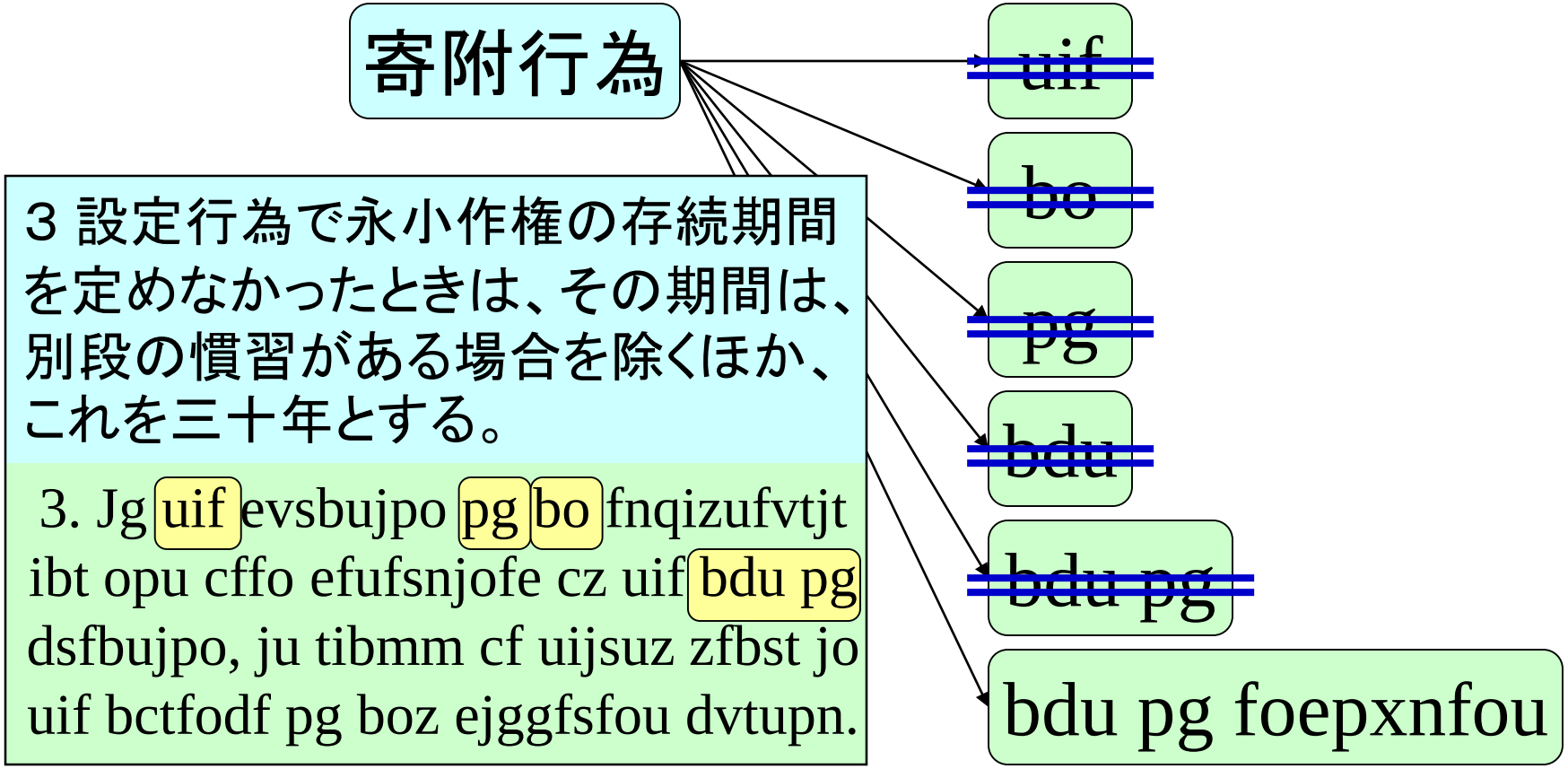
Base Idea

2. Pick up strings appeared in all of these articles

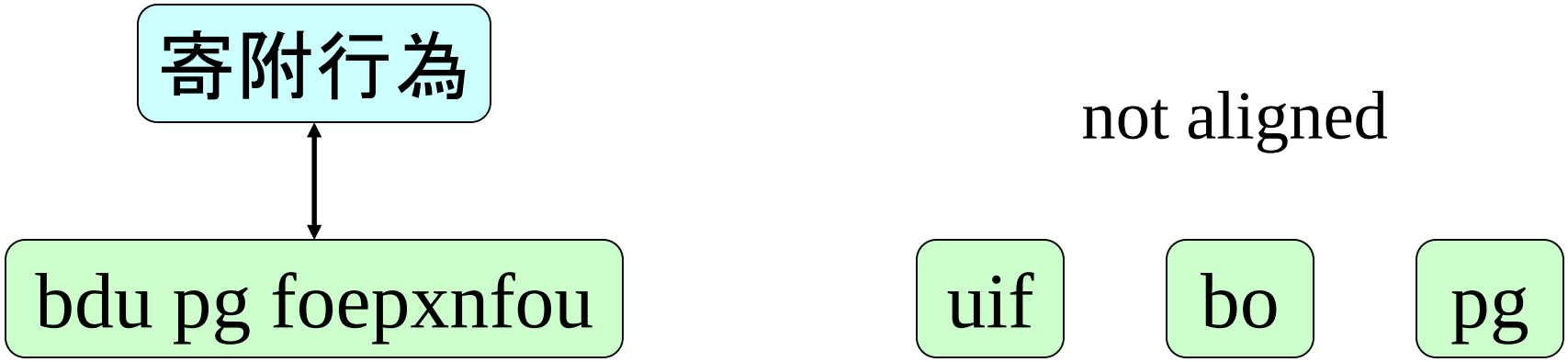
第三十九条 財団法人の設立者は、その設立を目的とする寄附行為で第三十七条第一号から第五号までに掲げた事項を定めなければならない。	Bsujdmf 39 1. Uif gpvoefs pg bo jodpsqpsbufe gpvoebujpo nvtu, jo uif bdu pg foepxnfou, nblf qspwjtjpo gps uif qbsujdvmbst jufnjafe jo Bsujdmf 37.
第四十一条 生前処分で寄附行為をするときは、贈与に関する規定を準用する。	Bsujdmf 41 1. Jg bo bdu pg foepxnfou jt epof cz b ejtqptjujpo joufs wjwpt, uif spwjtjpot sfmbujoh up hjgut tibmm bqqmz xjui ofdfttbsz npejgjdbujpot.
2. 遺言で寄附行為をするときは、遺贈に関する規定を準用する。	2. Jg bo bdu pg foepxnfou jt epof cz b xjmm, uif qspwjtjpot sfmbujoh up uftubnfoubsz cfrvftu tibmm bqqmz xjui ofdfttbsz npejgjdbujpot.

対訳候補

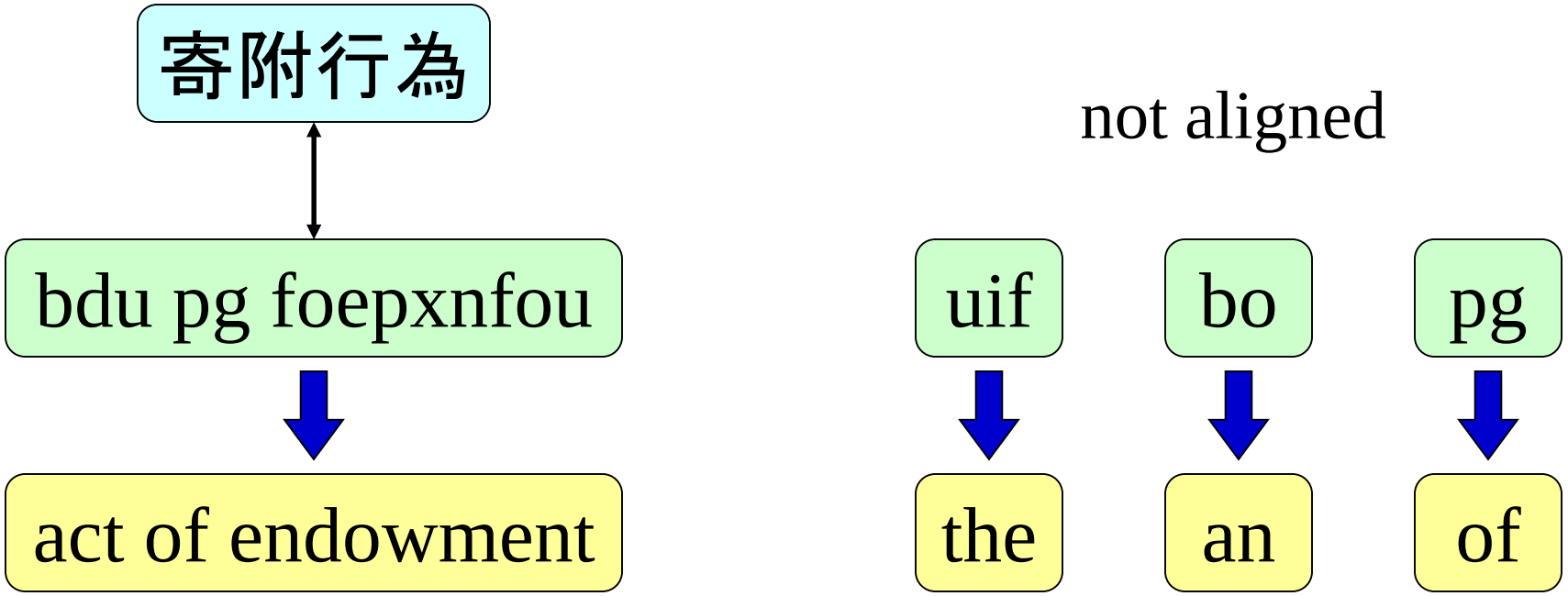
3. Eliminate the candidates which occur in the article whose source text doesn't contain "寄附行為"

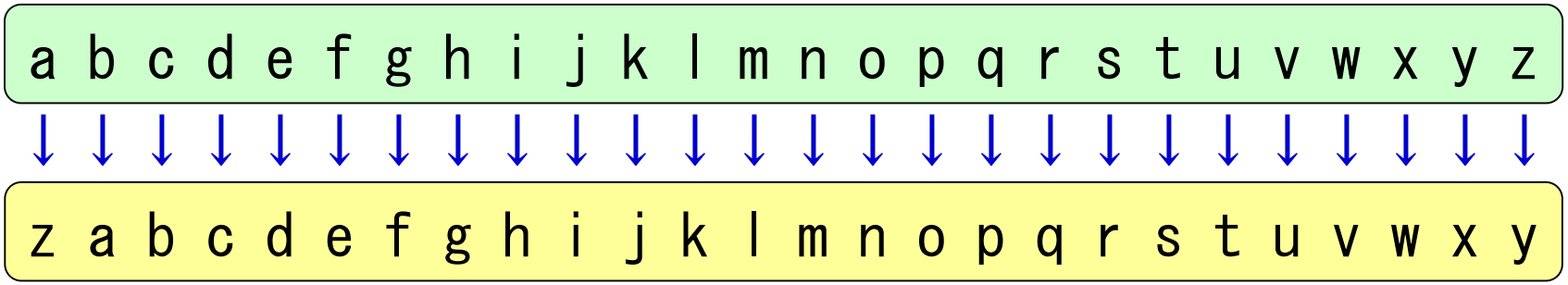


対訳語



架空の言語の答え


寄附行為
not aligned
bdu pg foepxnfou
uif
bo
pg
act of endowment
the
an
of


a b c d e f g h i j k l m n o p q r s t u v w x y z
z a b c d e f g h i j k l m n o p q r s t u v w x y

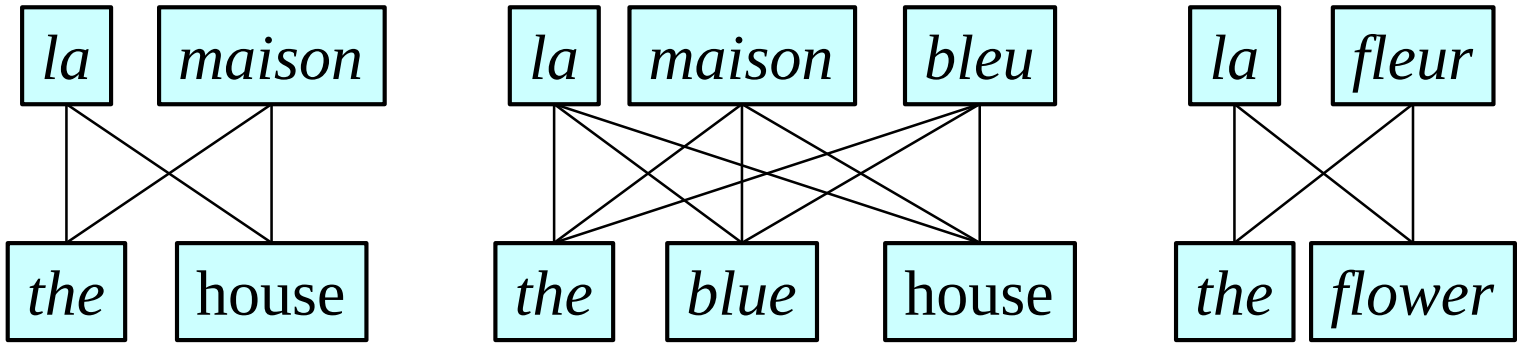
単語ベースのモデル

単語の翻訳確率 $t(e|f)$ に基づくモデル

- 外国語の単語 f が英語の単語 e に翻訳される確率

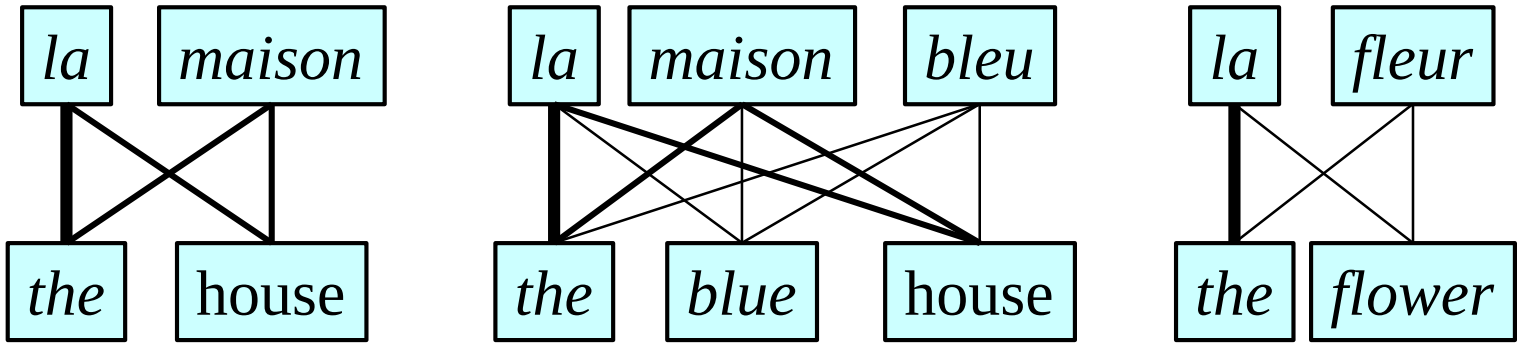
単語の翻訳確率の計算（1）

- EMアルゴリズムの利用



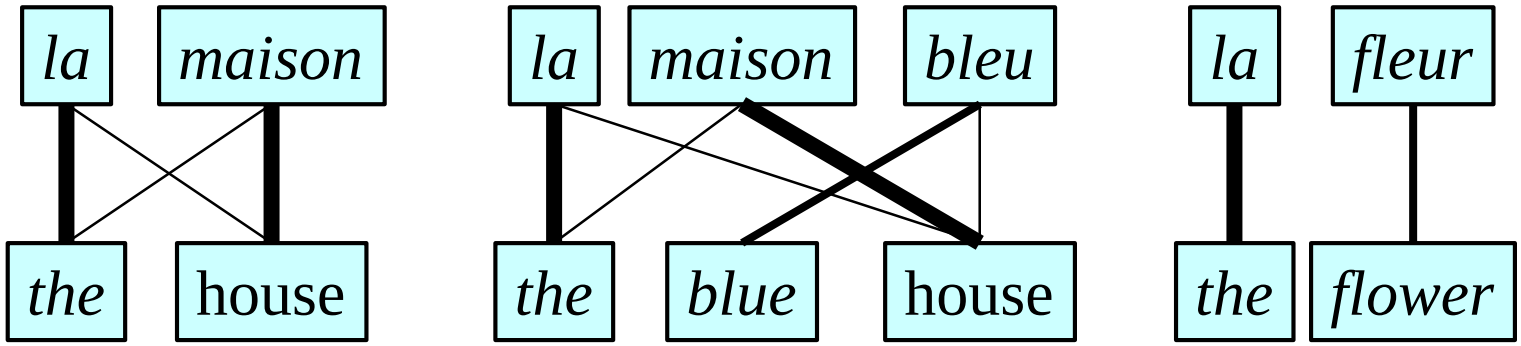
- 初期段階：すべての接続可能性が等しい
- *la* と *the* の接続が多い

単語の翻訳確率の計算(2)



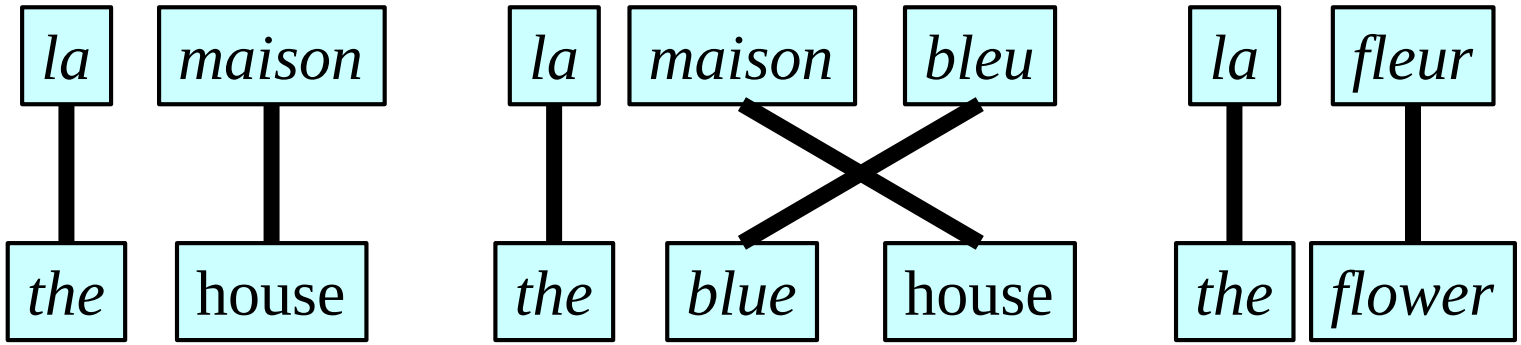
- 1回目の繰り返し
- la と the の接続可能性が増える

単語の翻訳確率の計算（3）



- 何回かの繰り返し
- *fleur* と *flower* の間などの接続可能性が増える（鳩の巣原理）

単語の翻訳確率の計算(4)



- 収束結果

IBMモデル1

- 単語の翻訳確率のみ考慮
- アラインメントは関数 a で表現

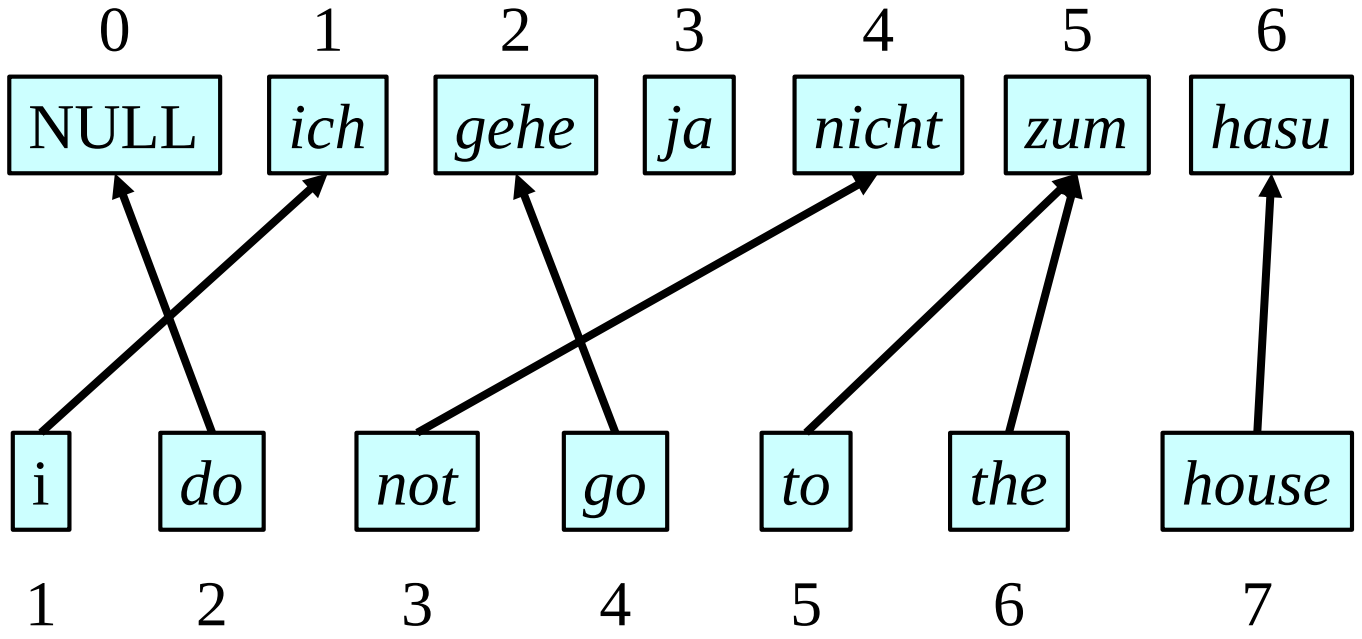
$$P(E, a|F) = \frac{\varepsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$

外国語文: $F = (f_1, \ldots, f_{l_f})$

英語文: $E = (e_1, \ldots, e_{l_e})$

ε: 正規化定数

アラインメント関数



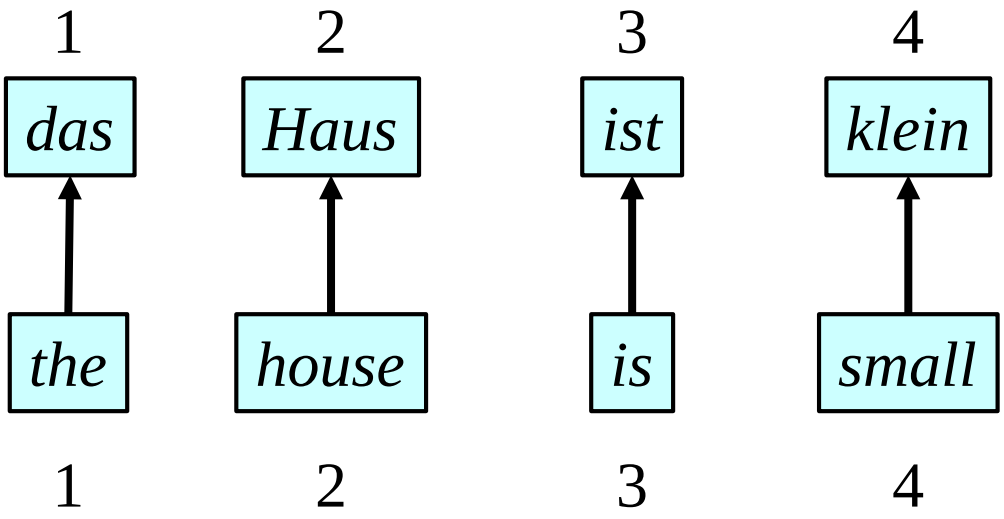
$a : \{1 \rightarrow 1, 2 \rightarrow 0, 3 \rightarrow 4, 4 \rightarrow 2, 5 \rightarrow 5, 6 \rightarrow 5, 7 \rightarrow 6\}$

IBMモデル1 続き

$$P(E|F) = \sum_{a} P(E, a|F)$$

$$P(E|F) = \frac{\varepsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} \sum_{i=0}^{l_f} t(e_j|f_i)$$

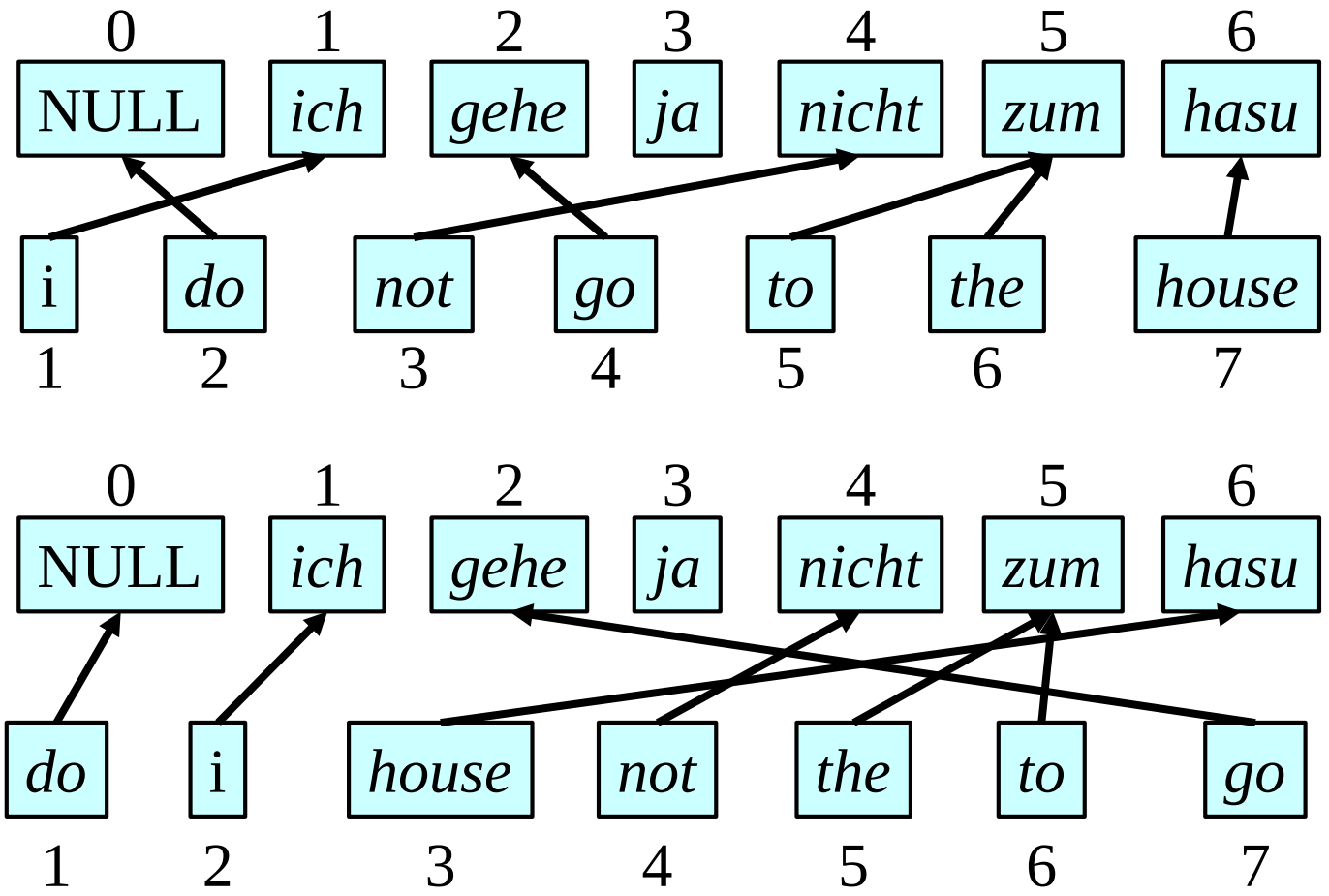
IBMモデル1の計算例



$$P(E, a|F) = \frac{\varepsilon}{5^4} \times t(the|das) \times t(house|Haus) \times t(is|ist) \times t(small|klein)$$

IBMモデル1の欠点

語順の違いを考慮しない



等確率

IBMモデル2

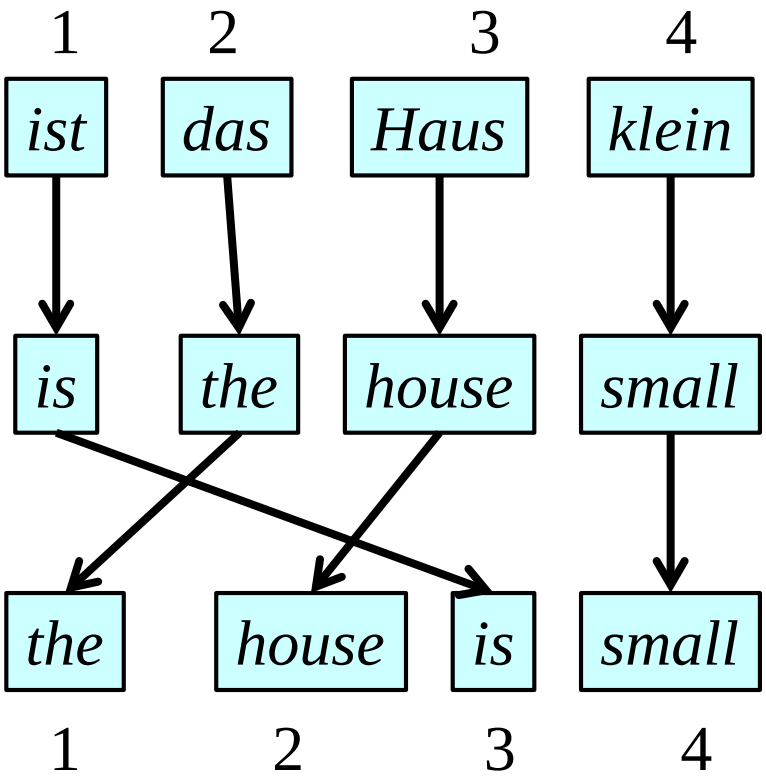
- アラインメントを確率で表現

$\alpha(i \mid j, l_e, l_f)$:

j番目の単語e_jがi番目の単語f_iに対応する確率

$$P(E|F) = \varepsilon \prod_{j=1}^{l_e} \sum_{i=0}^{l_f} t(e_j|f_i)\, \alpha(i \mid j, l_e, l_f)$$

アラインメントステップの導入



単語翻訳

アラインメント

IBMモデル2の欠点

単語の対応は1対1

IBMモデル3

- 産出力(fertility)を考慮

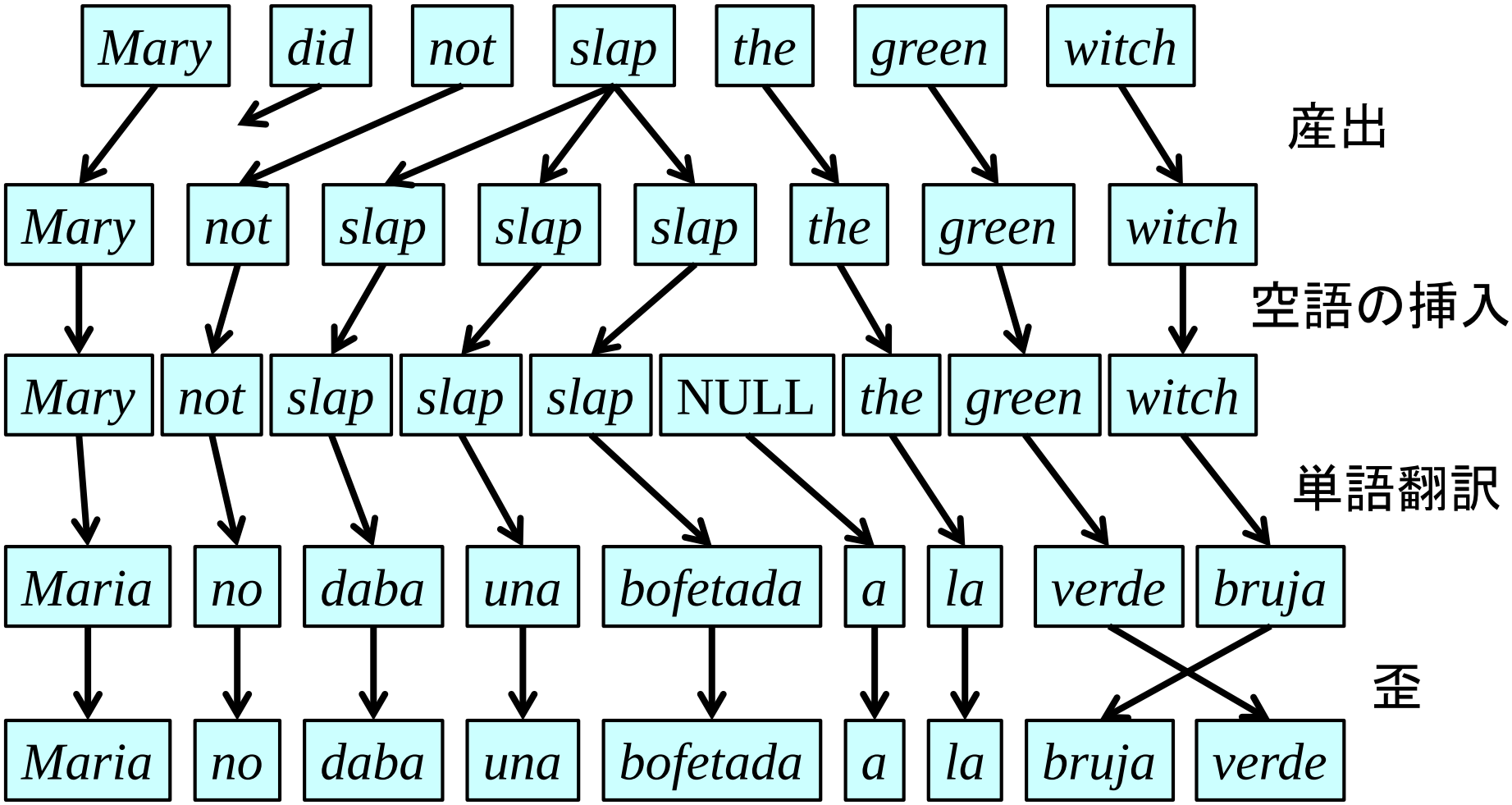
  $n(\varphi|e)$: e が φ 個の単語と対応する確率

- アラインメントの代わりに歪確率を考慮

$d(j \mid i, l_e, l_f)$:

  i番目の単語f_iがj番目の単語e_jに対応する確率
  アラインメントとは向きが逆

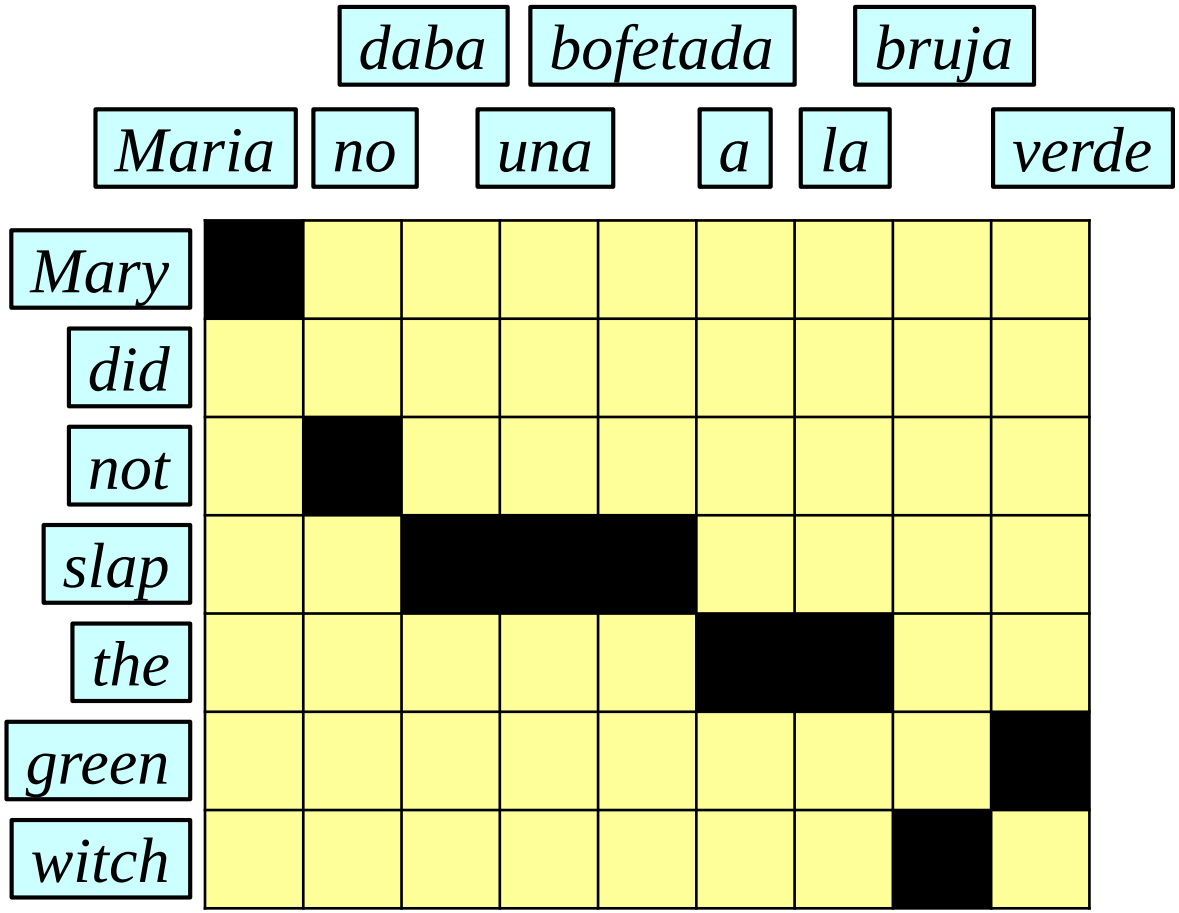
IBMモデル3の例


Mary did not slap the green witch
産出
Mary not slap slap slap the green witch
空語の挿入
Mary not slap slap slap NULL the green witch
単語翻訳
Maria no daba una bofetada a la verde bruja
歪
Maria no daba una bofetada a la bruja verde

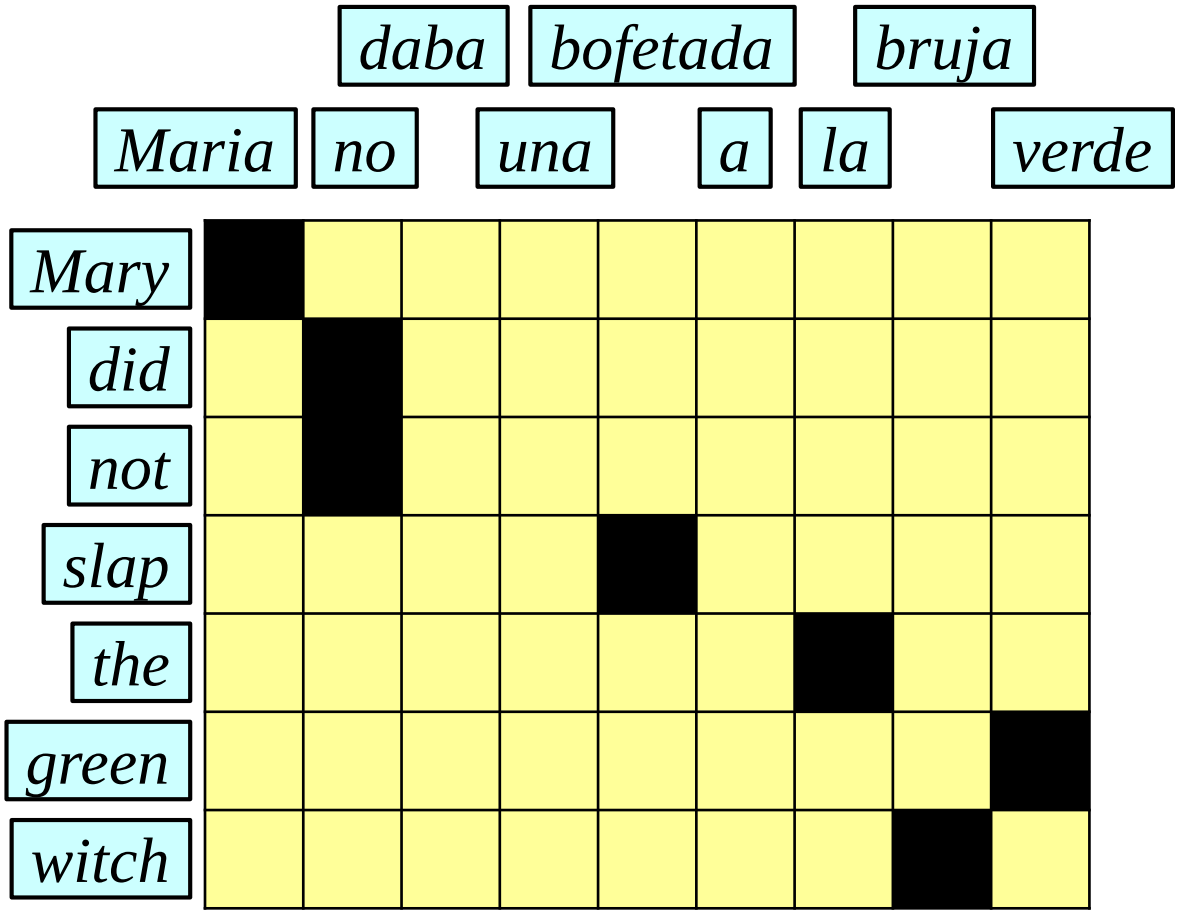
IBMモデル4及び5

- IBMモデル4
  - 歪確率を絶対位置から相対位置に変更
- IBMモデル5
  - 単語が同じ位置に配置されるのを修正

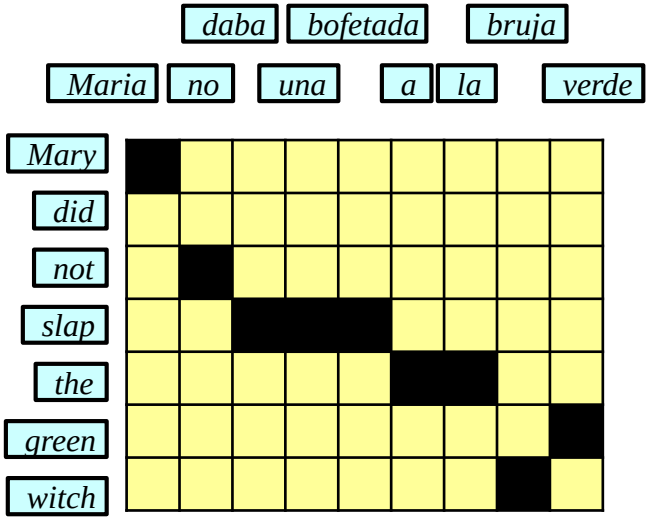
アラインメント　英語から西語

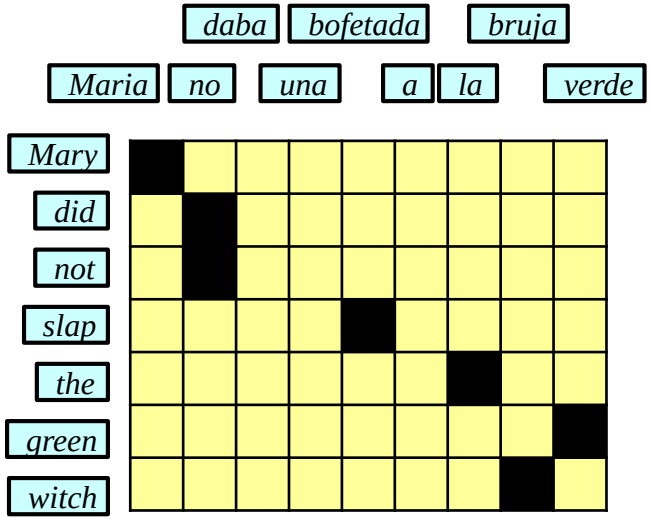


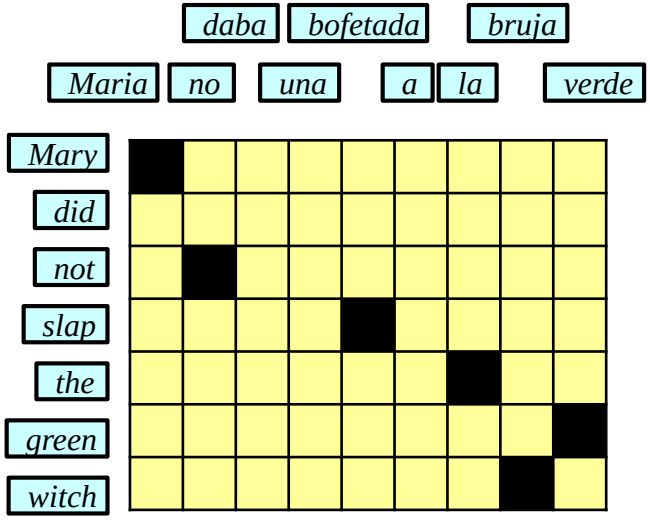
アラインメント　西語から英語



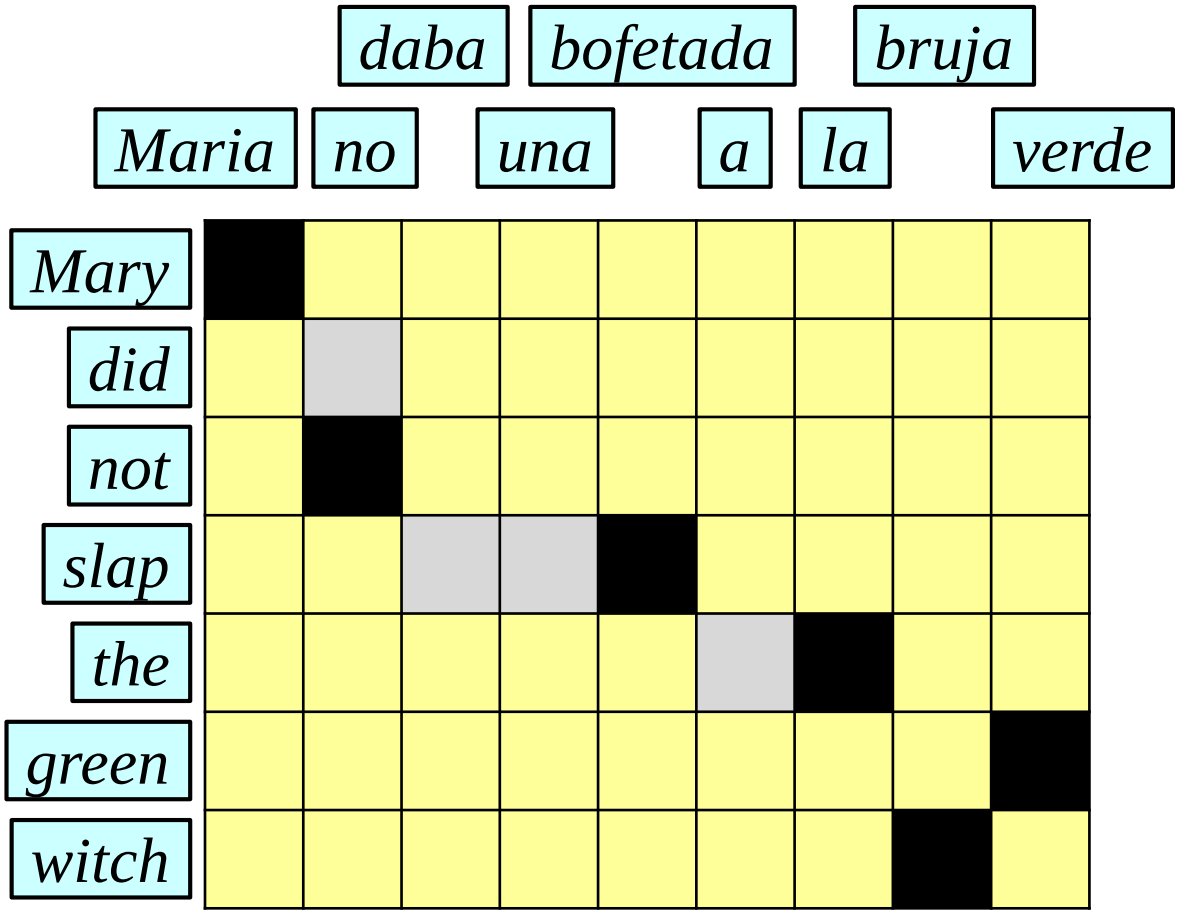
アラインメントの積


daba
bofetada
bruja
Maria
no
una
a
la
verde
Mary
did
not
slap
the
green
witch


daba
bofetada
bruja
Maria
no
una
a
la
verde
Mary
did
not
slap
the
green
witch


daba
bofetada
bruja
Maria
no
una
a
la
verde
Mary
did
not
slap
the
green
witch

アラインメントの改良

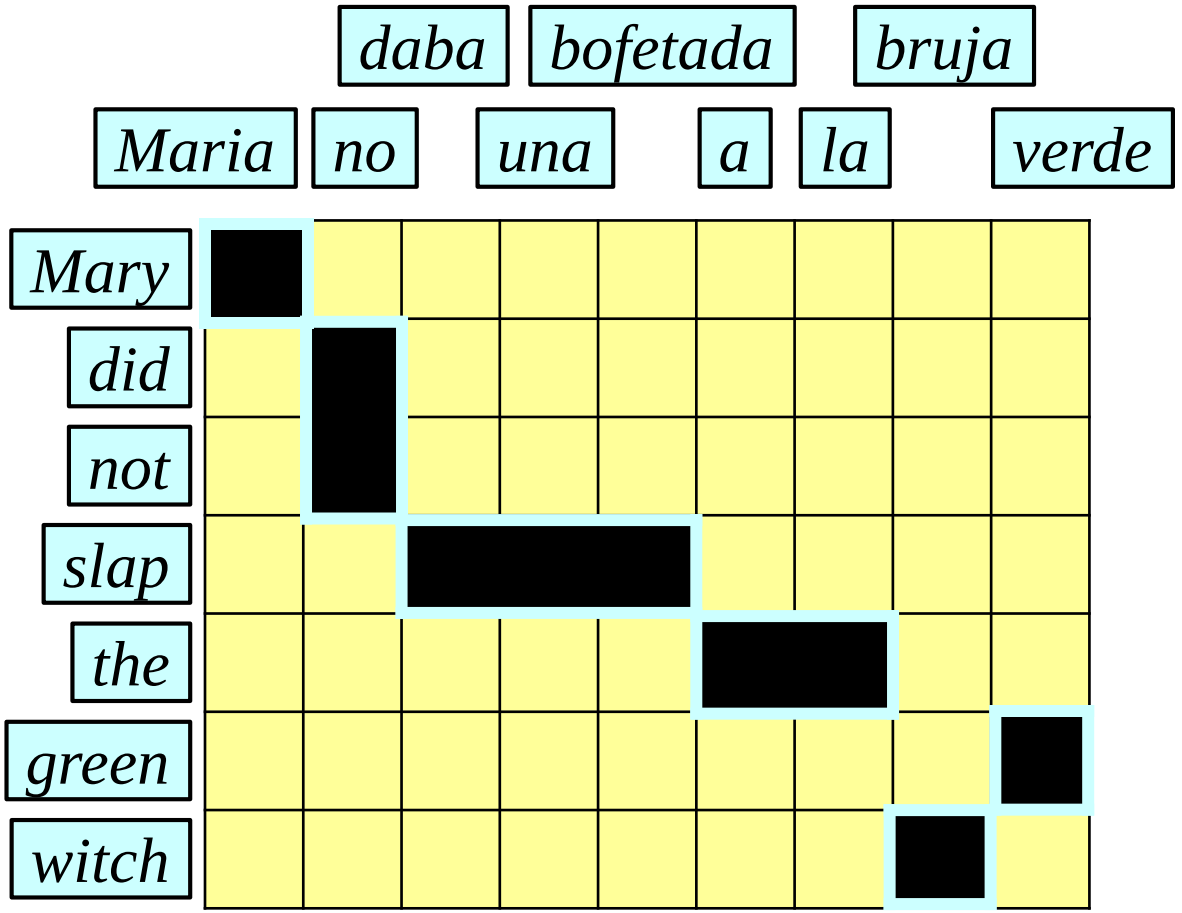

daba
bofetada
bruja
Maria
no
una
a
la
verde
Mary
did
not
slap
the
green
witch

句単位の翻訳

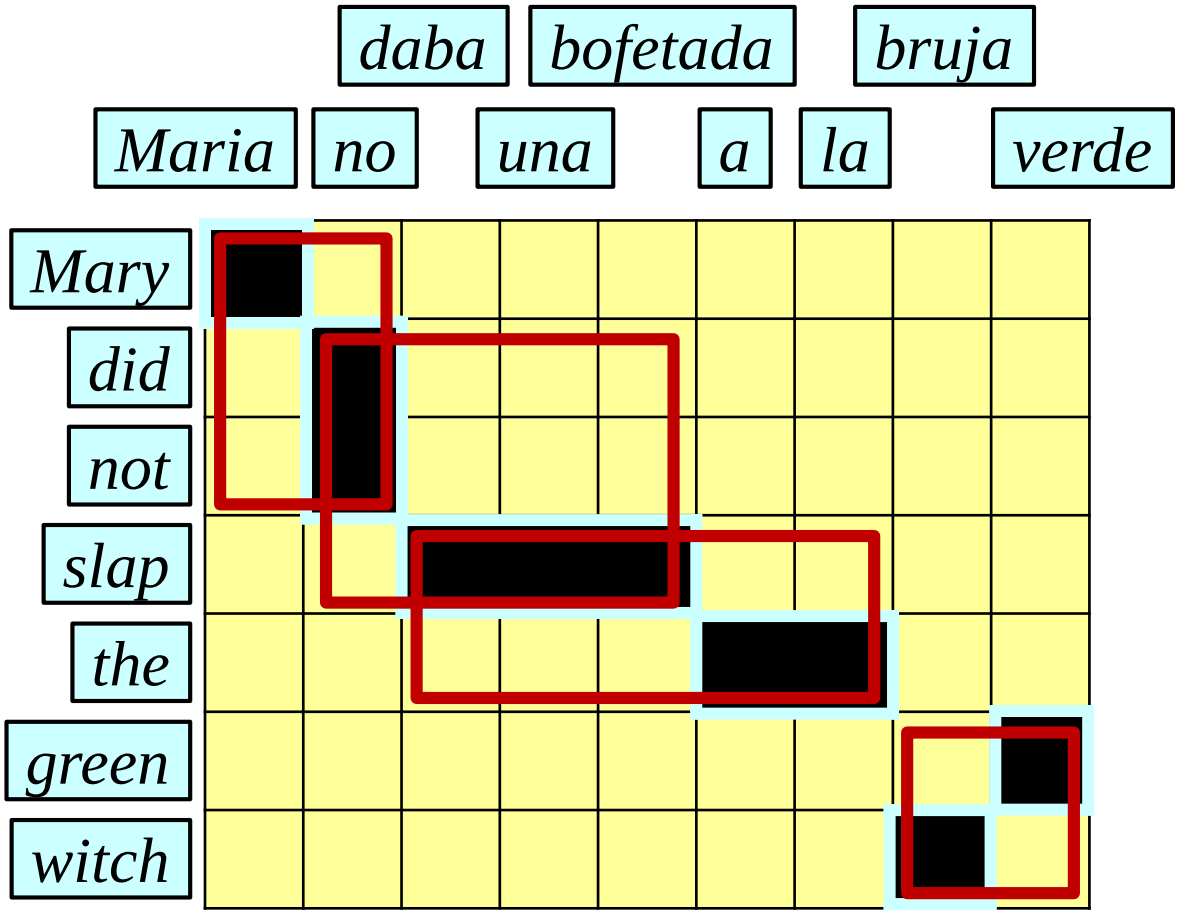
- 単語のまとまり単位での翻訳

$$\phi(\bar{f}|\bar{e}) = \frac{\mathrm{count}(\bar{e}, \bar{f})}{\sum_{\bar{f}_i} \mathrm{count}(\bar{e}, \bar{f}_i)}$$

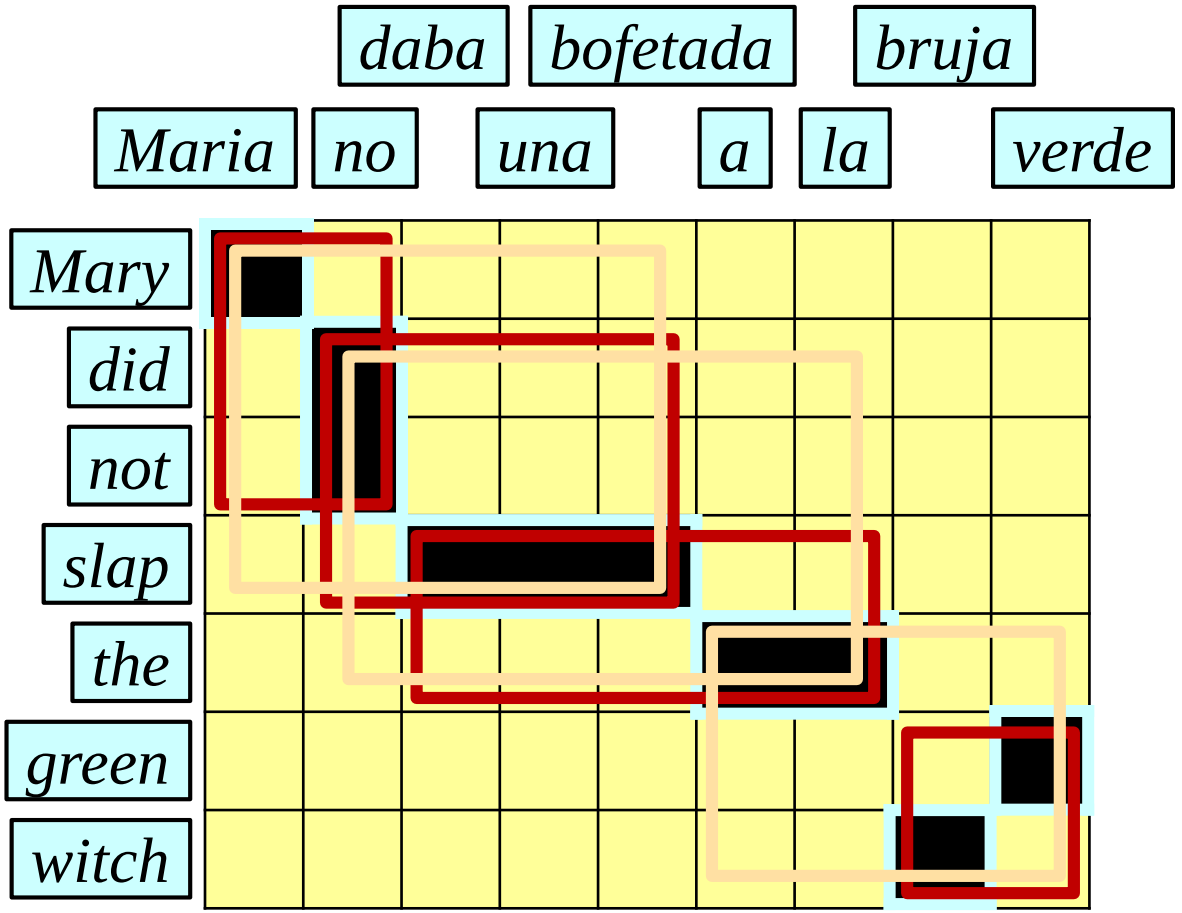
句の抽出


daba
bofetada
bruja
Maria
no
una
a
la
verde
Mary
did
not
slap
the
green
witch

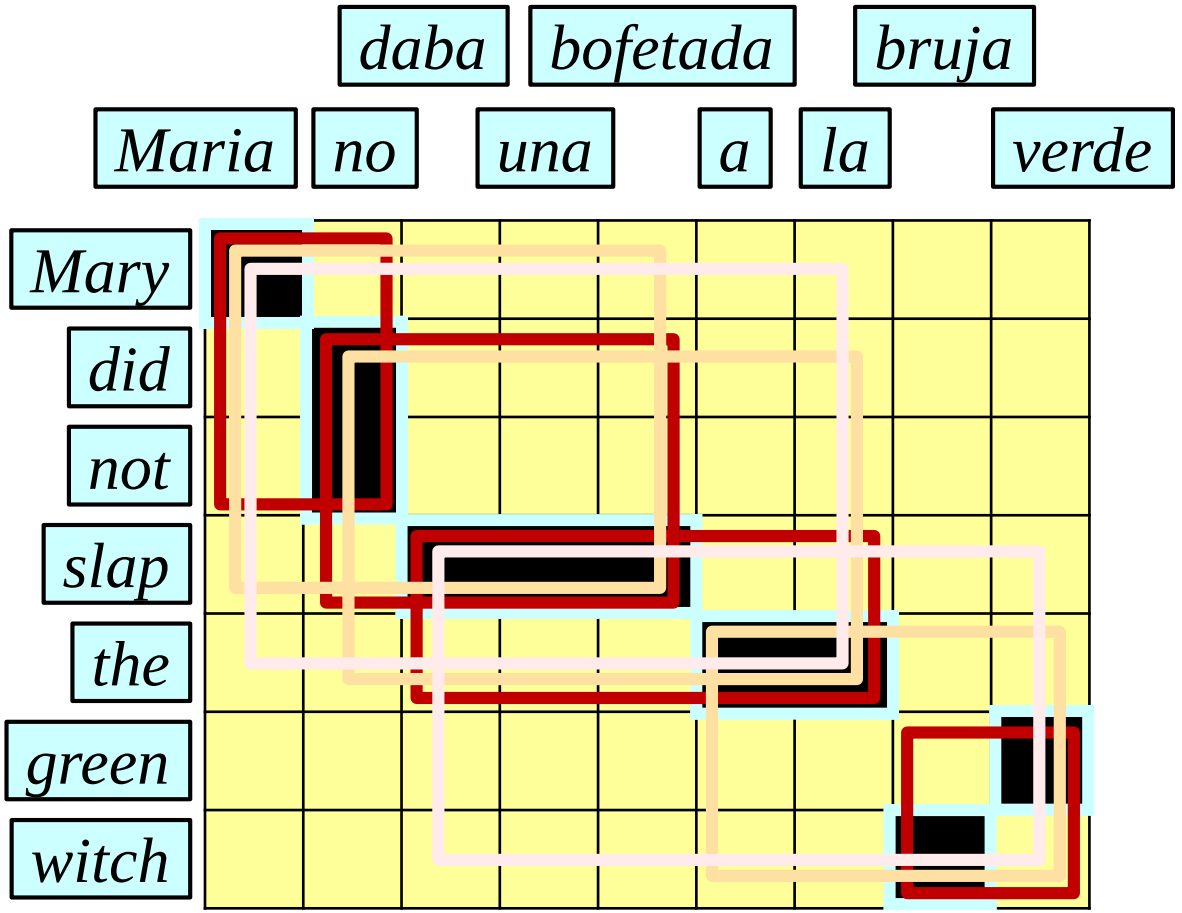
句の抽出


daba
bofetada
bruja
Maria
no
una
a
la
verde
Mary
did
not
slap
the
green
witch

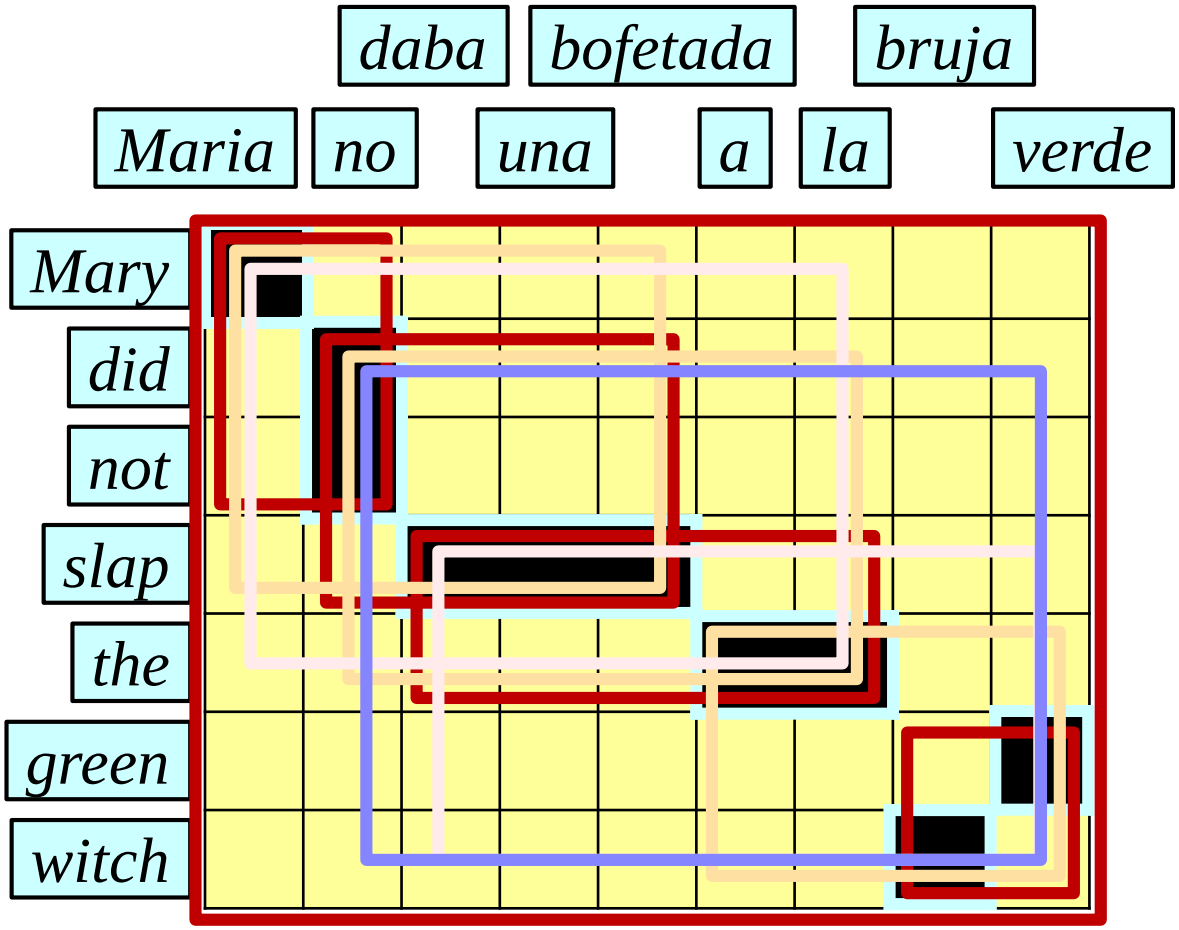
句の抽出


daba
bofetada
bruja
Maria
no
una
a
la
verde
Mary
did
not
slap
the
green
witch

句の抽出


daba
bofetada
bruja
Maria
no
una
a
la
verde
Mary
did
not
slap
the
green
witch

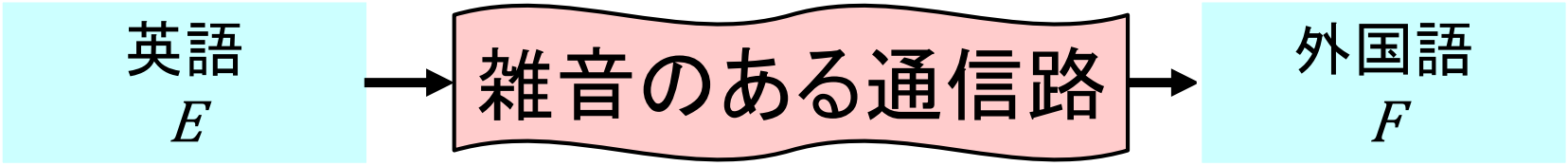
句の抽出


daba
bofetada
bruja
Maria
no
una
a
la
verde
Mary
did
not
slap
the
green
witch

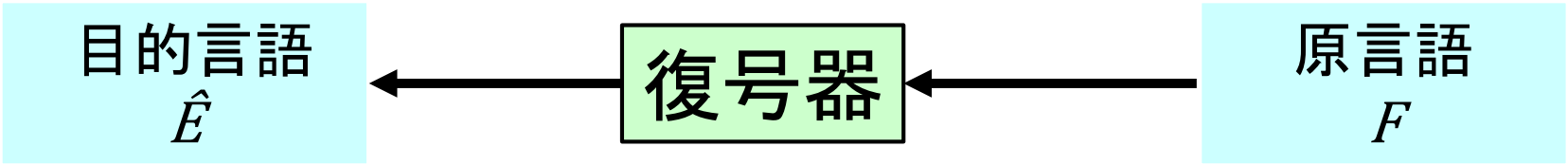
統計的機械翻訳への適用

外国語文から英語文への翻訳

- 英語文 E が雑音により外国語文 F となる



- 復号により英語文 $\hat{E}$ を推測



$$\hat{E} = \underset{E}{\operatorname{argmax}}\, P(E|F)$$

ベイズの定理

$$\hat{E} = \underset{E}{\operatorname{argmax}}\, P(E|F)$$

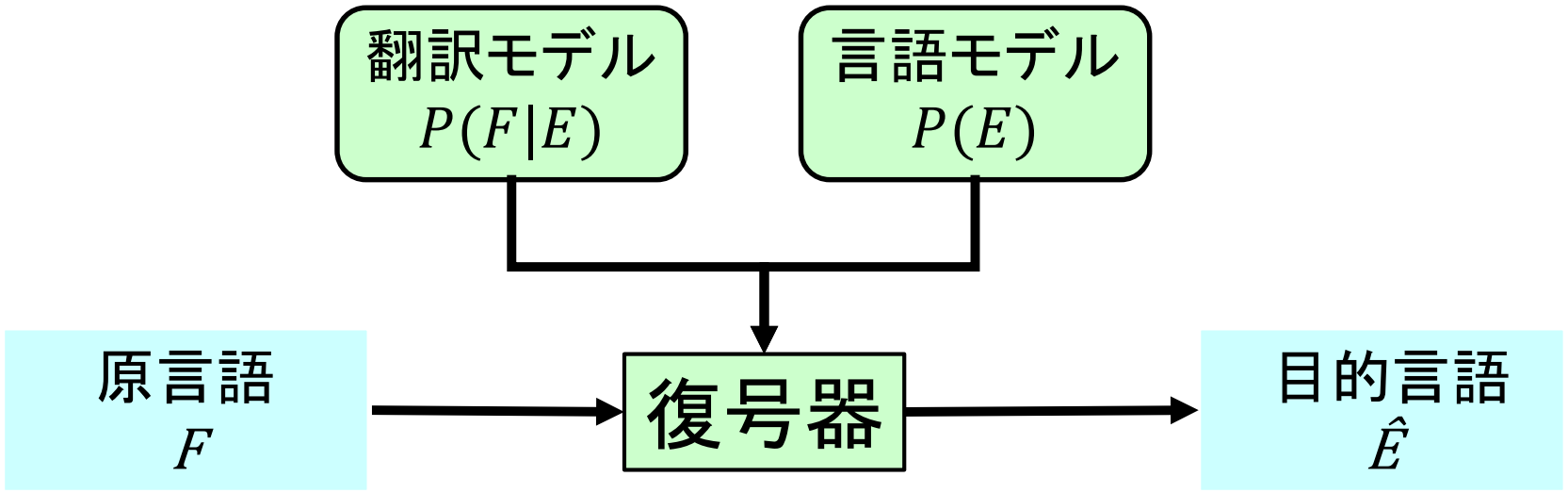
$$= \underset{E}{\operatorname{argmax}} \frac{P(F|E)P(E)}{P(F)}$$

$$= \underset{E}{\operatorname{argmax}}\, P(F|E)\, P(E)$$

翻訳モデル　言語モデル

統計的機械翻訳

$$\hat{E} = \underset{E}{\operatorname{argmax}}\, P(F|E)\, P(E)$$



言語モデル

- 言語としての確からしさ
  - nグラムモデル
    - スムージング（smoothing）

文法規則への対応

John has …

John have …

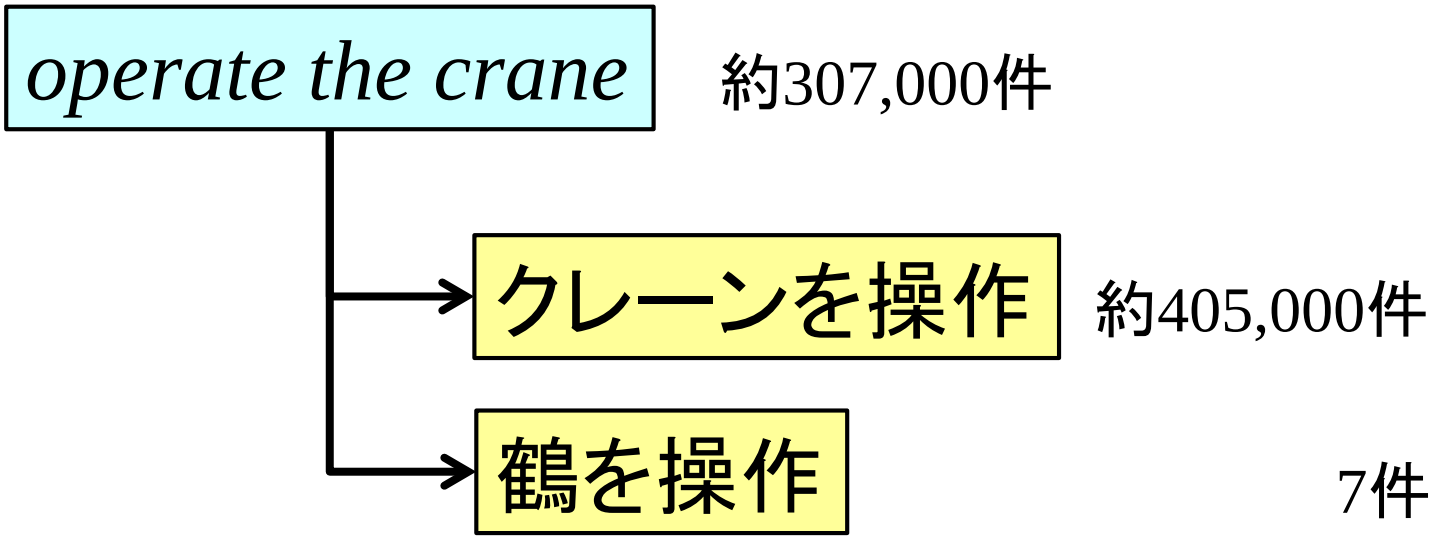
曖昧性解消

クレーンを操作

鶴を操作

言語モデルによる曖昧性解消

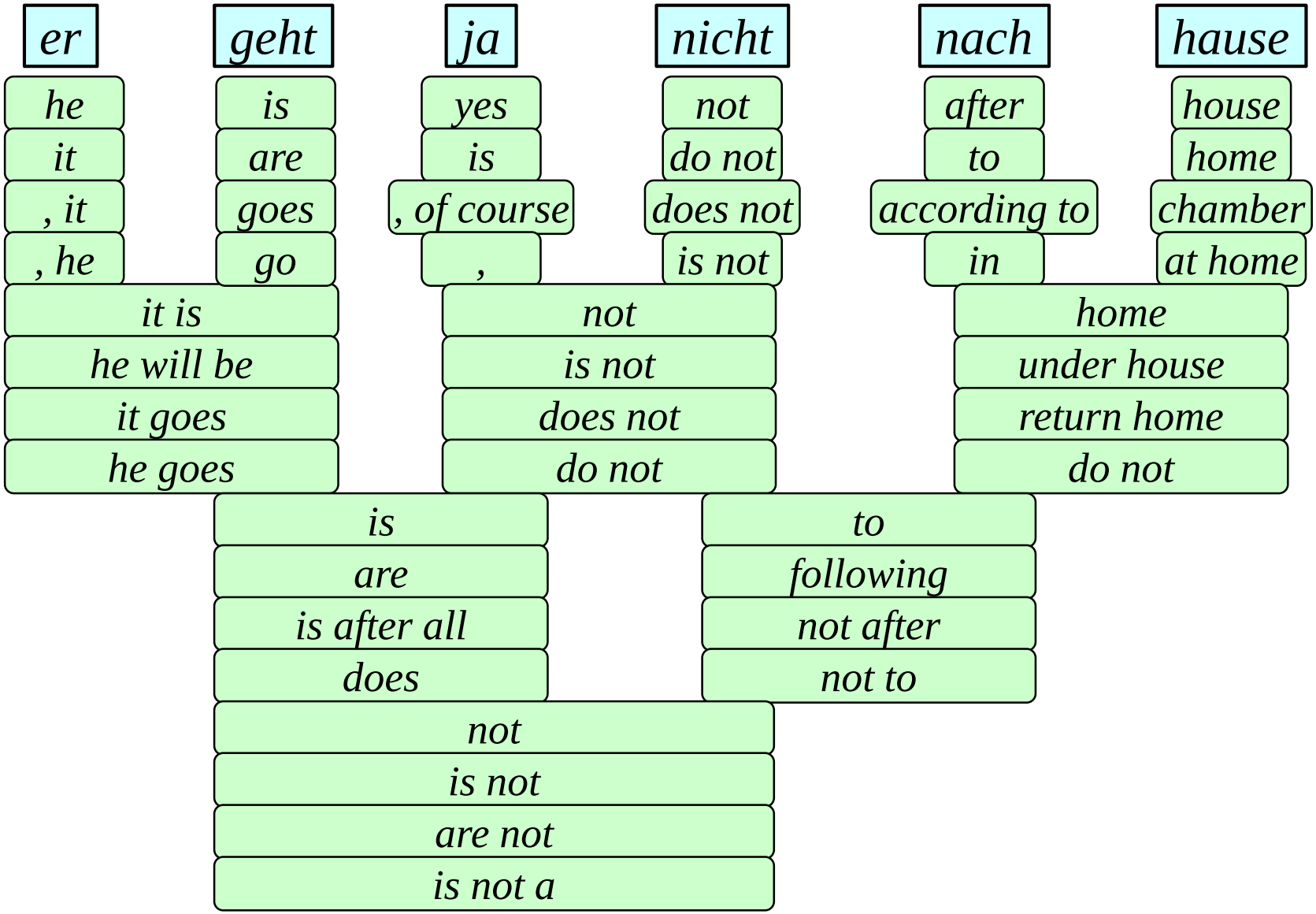
- コーパス中の出現回数から訳語を選択



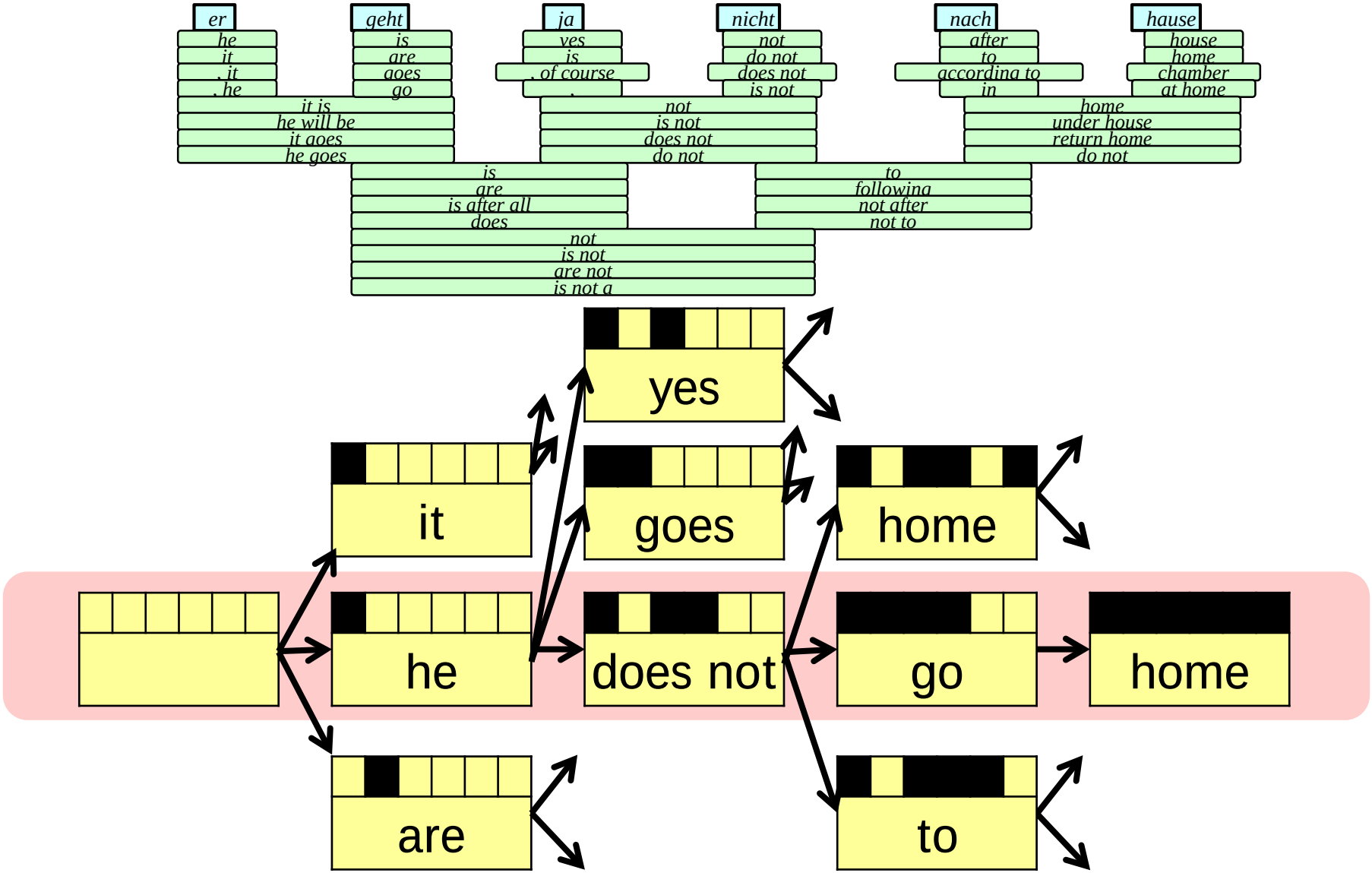
複合器(Decoder)

- 翻訳モデルと言語モデルを考慮
- サーチアルゴリズム
- 膨大な探索空間
  - どの範囲を探索するかはオプションで
    - 語の入れ替えの範囲

句に基づく翻訳の候補


er
geht
ja
nicht
nach
hause
he
it
, it
, he
is
are
goes
go
yes
is
, of course
,
not
do not
does not
is not
after
to
according to
in
house
home
chamber
at home
it is
he will be
it goes
he goes
not
is not
does not
do not
home
under house
return home
do not
is
are
is after all
does
to
following
not after
not to
not
is not
are not
is not a

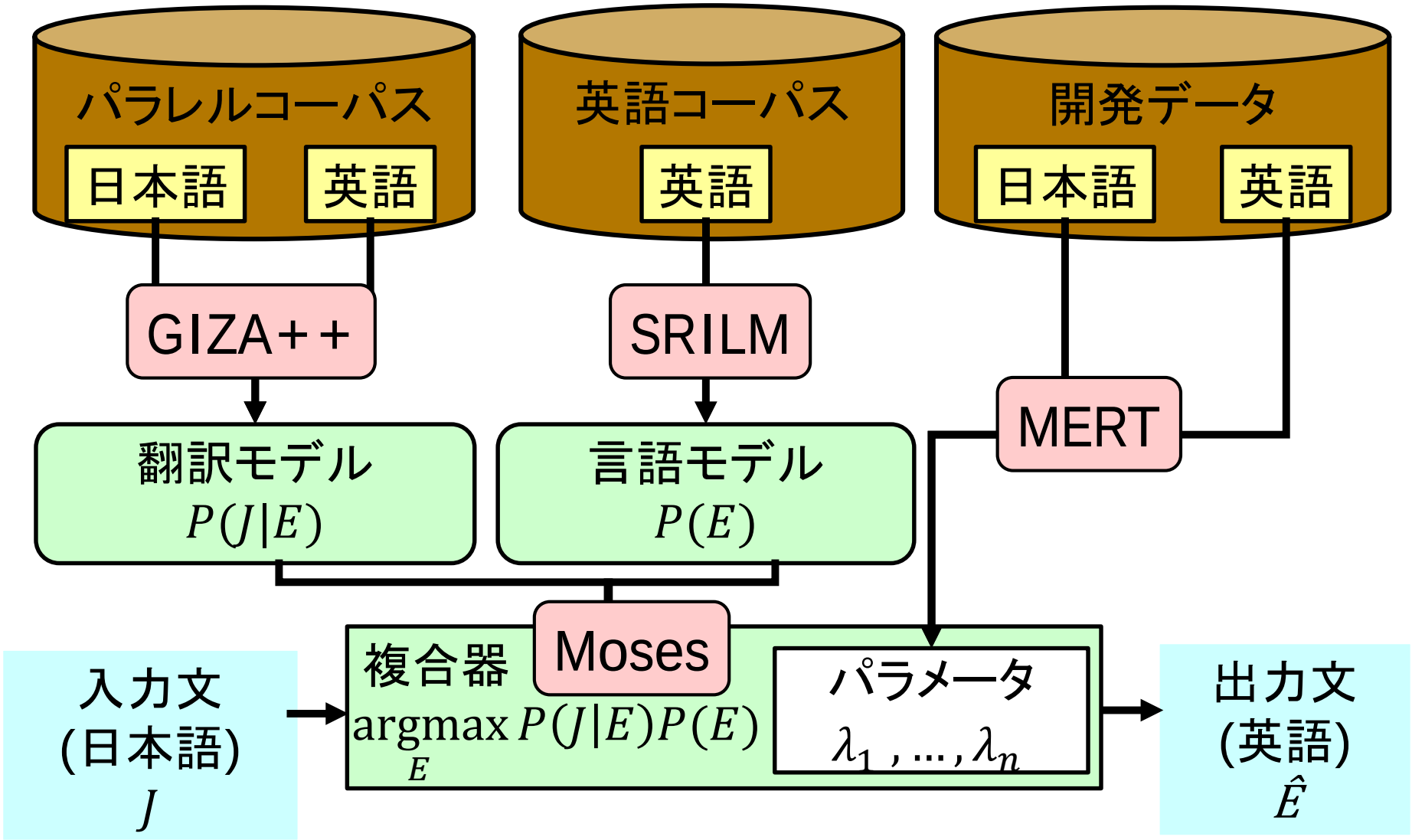
復号過程


er
he
it
, it
, he
it is
he will be
it goes
he goes
geht
is
are
goes
go
ja
yes
is
, of course
,
nicht
not
do not
does not
is not
not
is not
does not
do not
nach
after
to
according to
in
hause
house
home
chamber
at home
home
under house
return home
do not
is
are
is after all
does
not
is not
are not
is not a
to
following
not after
not to
yes
it
goes
home
he
does not
go
home
are
to

最小誤り率学習 (Och '03)

- 自動評価指標に合わせて
  複合器のパラメータを調整
  - 調整用のパラレルコーパス: 開発データ

統計的機械翻訳の構成


パラレルコーパス
日本語
英語
英語コーパス
英語
開発データ
日本語
英語
GIZA++
SRILM
MERT
翻訳モデル
$P(J|E)$
言語モデル
$P(E)$
Moses
入力文
(日本語)
J
複合器
$\underset{E}{\operatorname{argmax}} P(J|E)P(E)$
パラメータ
$\lambda_1, \ldots, \lambda_n$
出力文
(英語)
$\hat{E}$

翻訳の評価

- 人手による評価
  - 高コスト
    - 両言語の分かる専門家
  - 基準が一定でない
  - 量が多い
    - システムを変更するたびに別の翻訳結果

BLEU [Papineni et al. 2002]

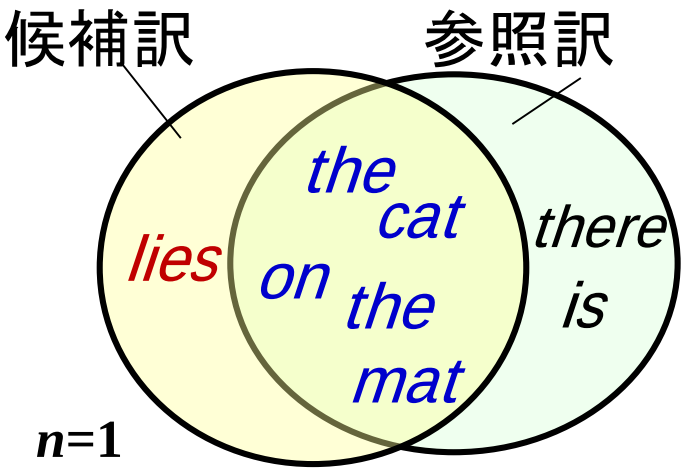
機械翻訳のための自動評価指標

- 正確性(adequacy)と流暢性(fluency)を評価
- 機械翻訳の出力（候補訳）と
  人間による翻訳（参照訳）を比較

候補訳: *The cat lies on the mat.*

参照訳

1. *The cat is one the mat.*
2. *There is the cat on the mat*



BLEU [Papineni et al. 2002]

$$p = \frac{\sum_{S \in Candidates} \sum_{w \in S} Count_{clip}(w)}{\sum_{S \in Candidates} \sum_{w \in S} Count(w)}$$

候補訳と参照訳に共起した回数

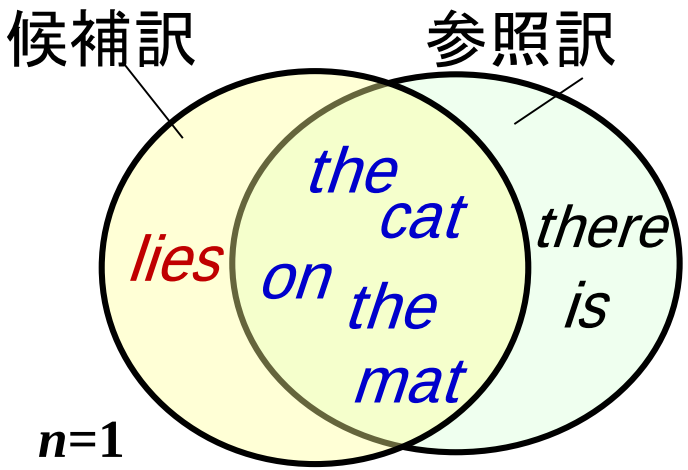
候補訳に出現した回数

候補訳: *The cat lies on the mat.*

参照訳

1. *The cat is one the mat.*
2. *There is the cat on the mat*

$$p_1 = \frac{5}{6}$$



BLEU (cont'd)

BLEU Score

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

BP: brevity penalty

N = 4

$w_n = 1/N$

- 小さな n: 正確性を評価
- 大きな n: 流暢性を評価

まとめ

- 統計的機械翻訳
  - コーパスからの学習
  - 言語に依存しない
  - コーパスの量が重要
  - 語順が似ている言語間では高性能

オマケ

Garden path sentence

The girl told the story cried.

その物語を聞いた少女は泣いた。

The old man the boat.

老人達はボートに人を配置した。

The raft floated down the river sank.

川を浮きながら下ったいかだは沈んだ。