言語学的に妥当な CCGツリーバンク構築の試み

第37回 人工知能学会全国大会　2023/06/07

富田朝[*1]　谷中瞳[*2]　戸次大介[*1]

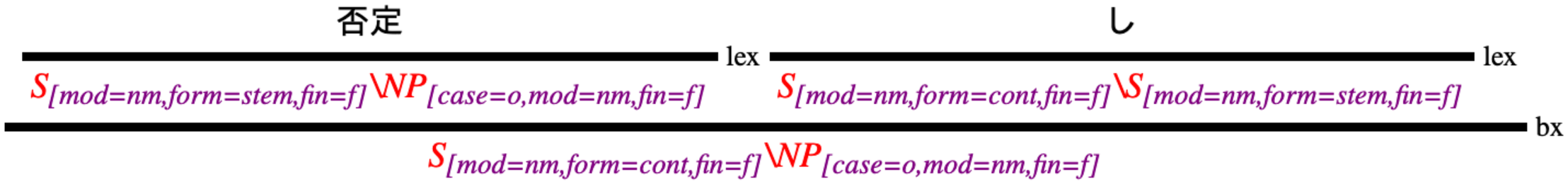
[*1] お茶の水女子大学　[*2] 東京大学

ツリーバンク(treebank)

- 各文に**統語構造**が付与されている**コーパス**

  → 語同士がどのような順序や規則で結合しているのかを表す構造

例：日本語CCGbank [Uematsu+ 2013]



組み合わせ範疇文法 (Combinatory Categorial Grammar; CCG) [Steedman 1996][Steedman 2001]

辞書

語とその統語情報や意味情報を関連付ける

Keats ⊢ NP
eats ⊢ $(S \setminus NP) / NP$
apples ⊢ NP

＋

組み合わせ規則

例: 関数適用規則
関数合成規則

背景と目的

CCGパーザ：自然言語を入力として受け取り、CCG木を出力する

- **depccg**[Yoshikawa+ 2017], **jigg**[Noji and Miyao 2016]などがある
- 学習・評価データとして**CCGツリーバンク**を利用する
- 妥当性が**CCGツリーバンクの妥当性に依存**する

言語学的に妥当なCCGツリーバンクの構築が必要
→標準的な日本語CCGの議論がされているBekki (2010)に基づく

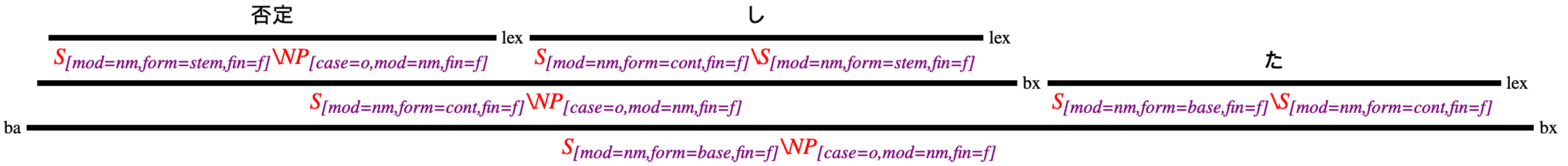
日本語CCGbank [Uematsu+ 2013]

京都大学
テキストコーパス
[Kawahara+ 2002]

NAIST
テキストコーパス
[Iida+ 2007]

「と」
コーパス
[Hanaoka+ 2010]

CCG木へ自動変換して構築



日本語CCGbank [Uematsu+ 2013]

課題：受身・使役の構文に対して誤った分析がなされていることが指摘されている [Bekki and Yanaka 2023]

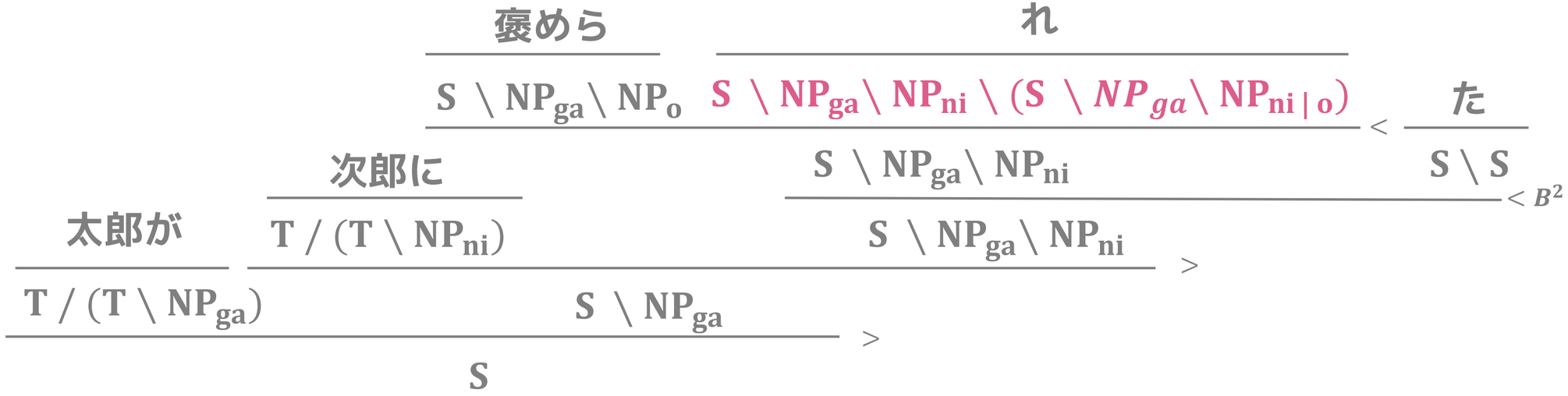
日本語CCGbankの分析

$$\dfrac{\dfrac{\text{太郎が}}{\mathbf{NP_{ga}}}\quad\dfrac{\dfrac{\text{次郎に}}{\mathbf{NP_{ni}}}\quad\dfrac{\dfrac{\dfrac{\text{褒めら}}{\mathbf{S\backslash NP_{ga}\backslash NP_{ni}}}\quad\dfrac{\text{れ}}{\mathbf{S\backslash S}}}{\mathbf{S\backslash NP_{ga}\backslash NP_{ni}}}<B^2\quad\dfrac{\text{た}}{\mathbf{S\backslash S}}}{\mathbf{S\backslash NP_{ga}\backslash NP_{ni}}}<B^2}{\mathbf{S\backslash NP_{ga}}}<}{\mathbf{S}}<$$

日本語CCGbank [Uematsu+ 2013]

課題：受身・使役の構文に対して誤った分析がなされていることが指摘されている [Bekki and Yanaka 2023]

言語学的に妥当なCCG木



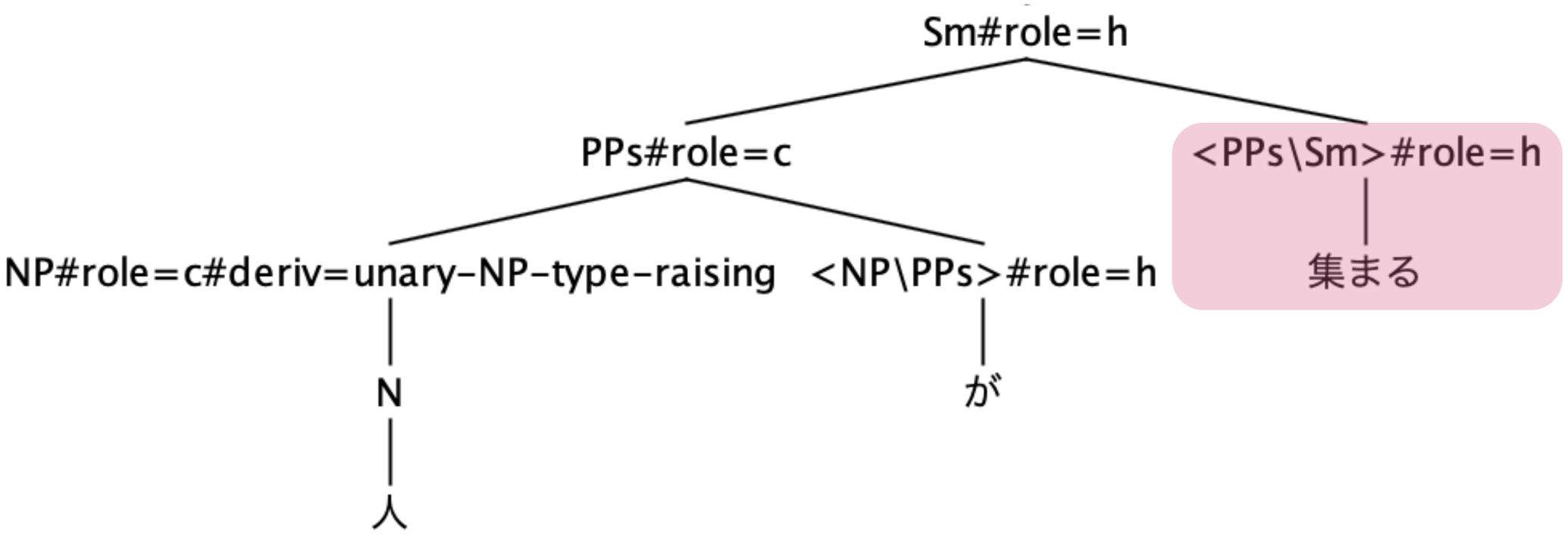
ABCツリーバンク [Kubota+ 2019]

- けやきツリーバンク[Butler 2012]を**ABC文法**へ変換して構築
  （ABC文法：関数適用規則＋関数合成規則）

- CCGやTLG[Morrill 1994; Moortgat 1997; Kubota and Levine 2015]のツリーバンクに変換できる

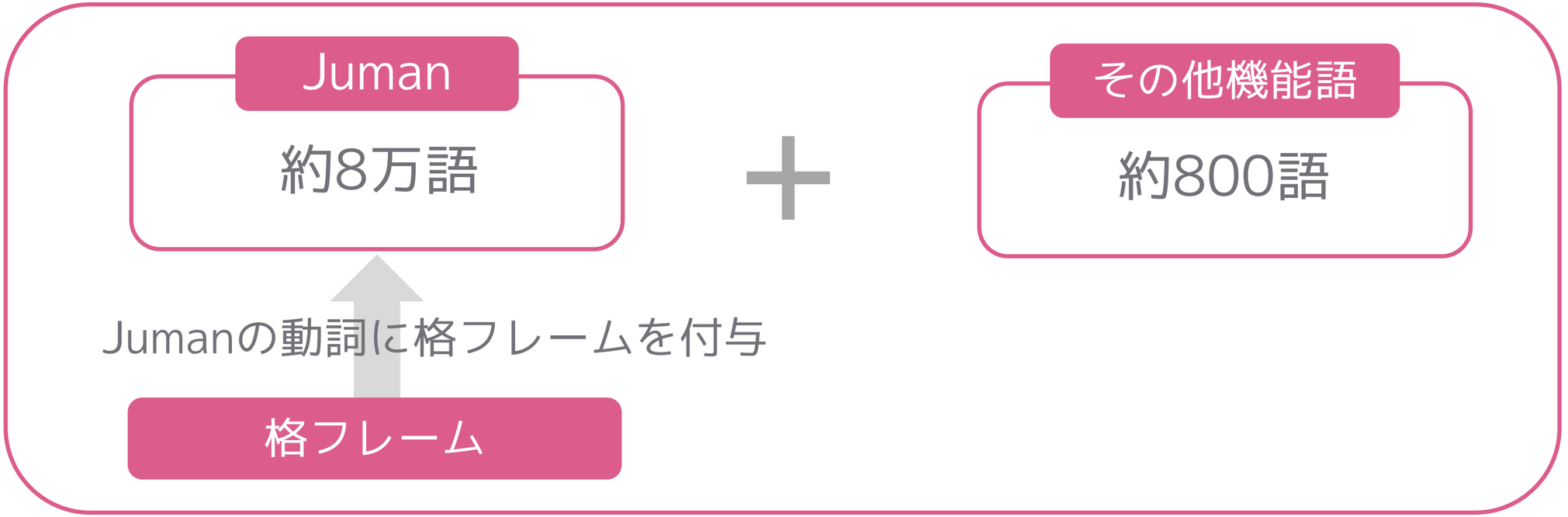
ABCツリーバンク [Kubota+ 2019]

特徴：項構造などが人手によって記述されている
課題：**活用の種別**などの統語情報が含まれない



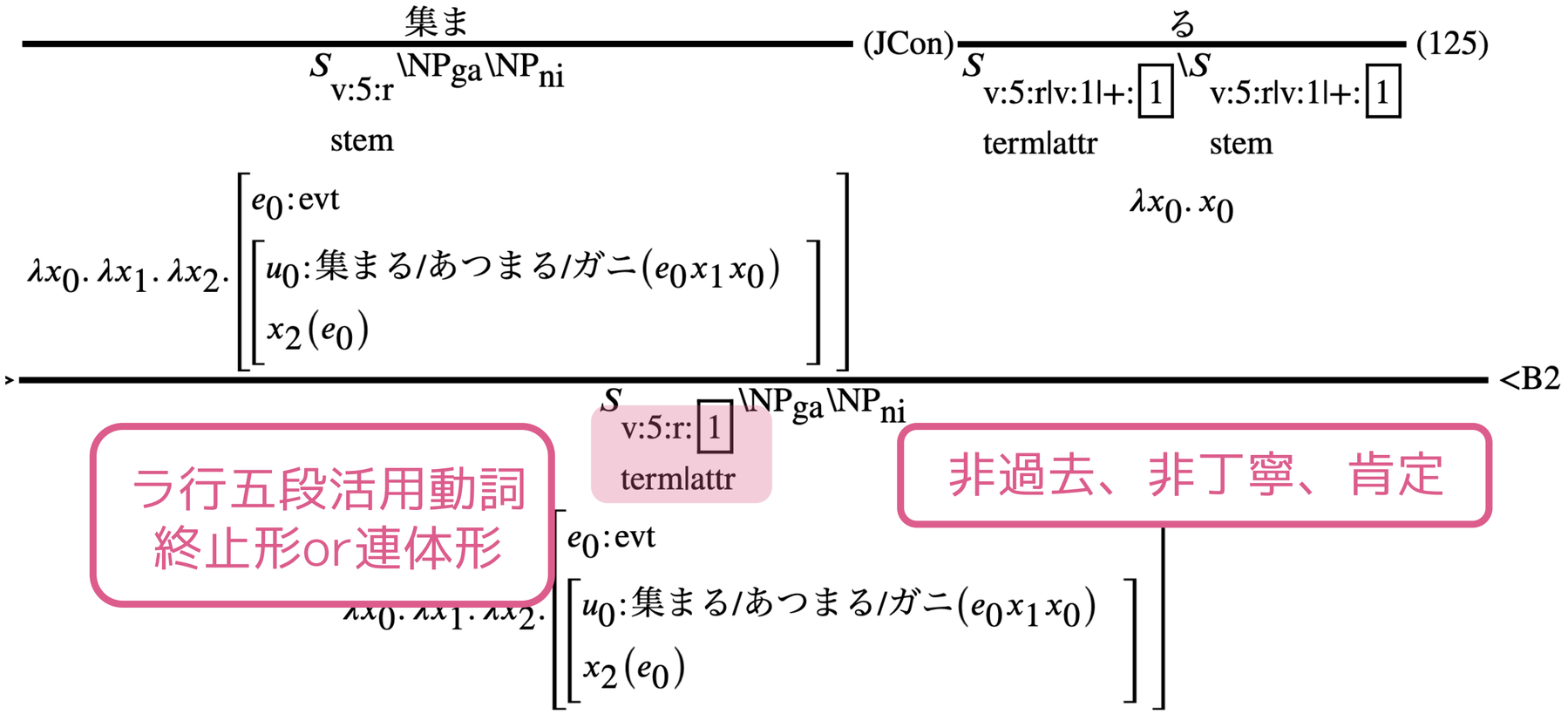
lightblue [Bekki and Kawazoe 2016]

- 格フレームを元に作られた**辞書**とCCGの組合せ規則に基づいてCCG統語解析を行う



lightblue [Bekki and Kawazoe 2016]

特徴：詳細な統語情報を含むCCG木を出力する



lightblue [Bekki and Kawazoe 2016]

特徴：詳細な統語素性を含むCCG木を出力する

課題：項構造に関する誤りが多い

- 与えられた文脈においては自然ではない用言
- 用言の格フレームに誤りが含まれる

研究目的と提案手法

研究目的

言語学的に妥当で**詳細な統語情報**を有する
日本語CCG ツリーバンクの構築

研究目的と提案手法

研究目的

言語学的に妥当で**詳細な統語情報**を有する
日本語CCG ツリーバンクの構築

提案手法

ABCツリーバンク

人手によって**項構造が正確**に記述されている

＋

lightblue

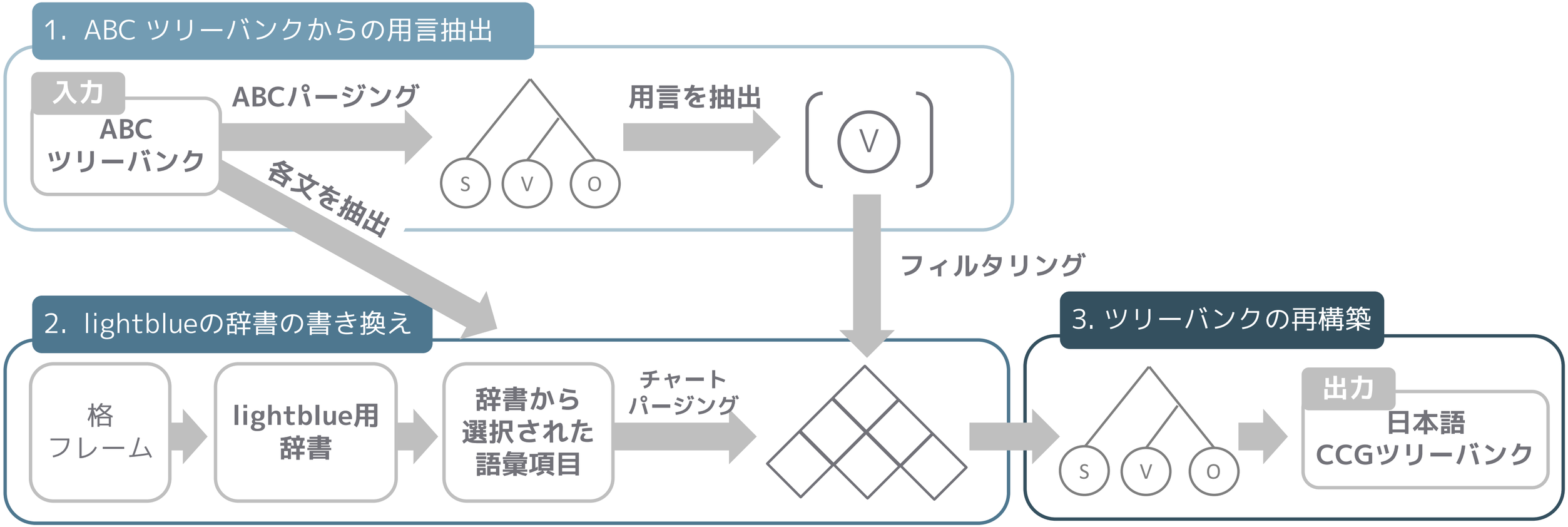
詳細な統語情報を与えることができる

ABCツリーバンクのリフォージング

リフォージング：ツリーバンクを分解し、再構築する手法

1. ABCツリーバンクからの用言抽出
2. lightblueの辞書の書き換え
3. ツリーバンクの再構築

ABCツリーバンクのリフォージング

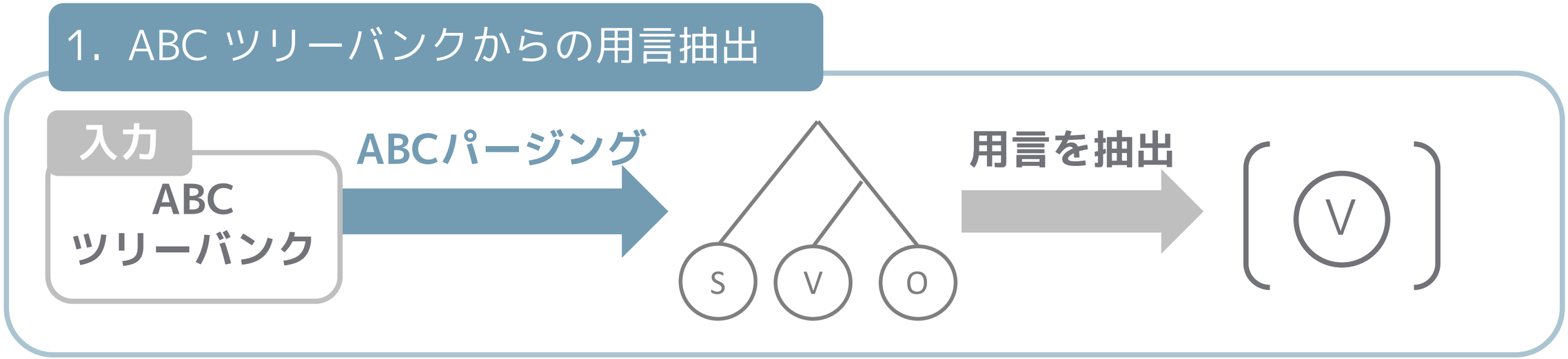


リフォージング – ABCツリーバンクからの用言抽出

1. ABCツリーバンクからの用言抽出
2. lightblueの辞書の書き換え
3. ツリーバンクの再構築

リフォージング – ABCツリーバンクからの用言抽出

- ABCツリーバンク（テキストデータ）を木構造データで扱えるようにparseする

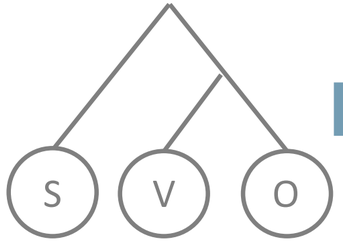


リフォージング – ABCツリーバンクからの用言抽出

1. ABC ツリーバンクからの用言抽出

入力
ABC
ツリーバンク

ABCパージング



用言を抽出

例

入力
「嫌いな人もいる」

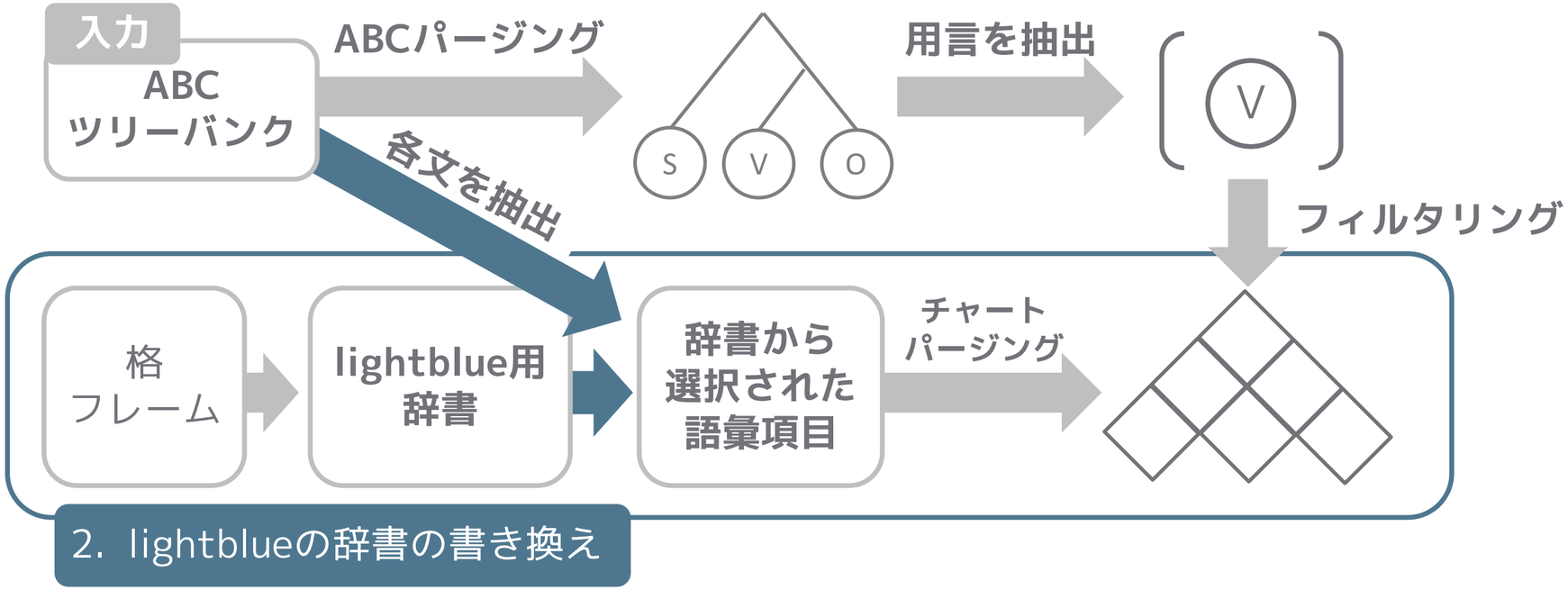
出力

[("嫌い", PPs \ Srel , 0 , 1),
表層形　統語範疇　開始位置　終了位置
(“いる”, PPs \ Sm , 5 , 6)]

リフォージング – lightblueの辞書の書き換え

1. ABCツリーバンクからの用言抽出
2. lightblueの辞書の書き換え
3. ツリーバンクの再構築

リフォージング – lightblueの辞書の書き換え



例

入力

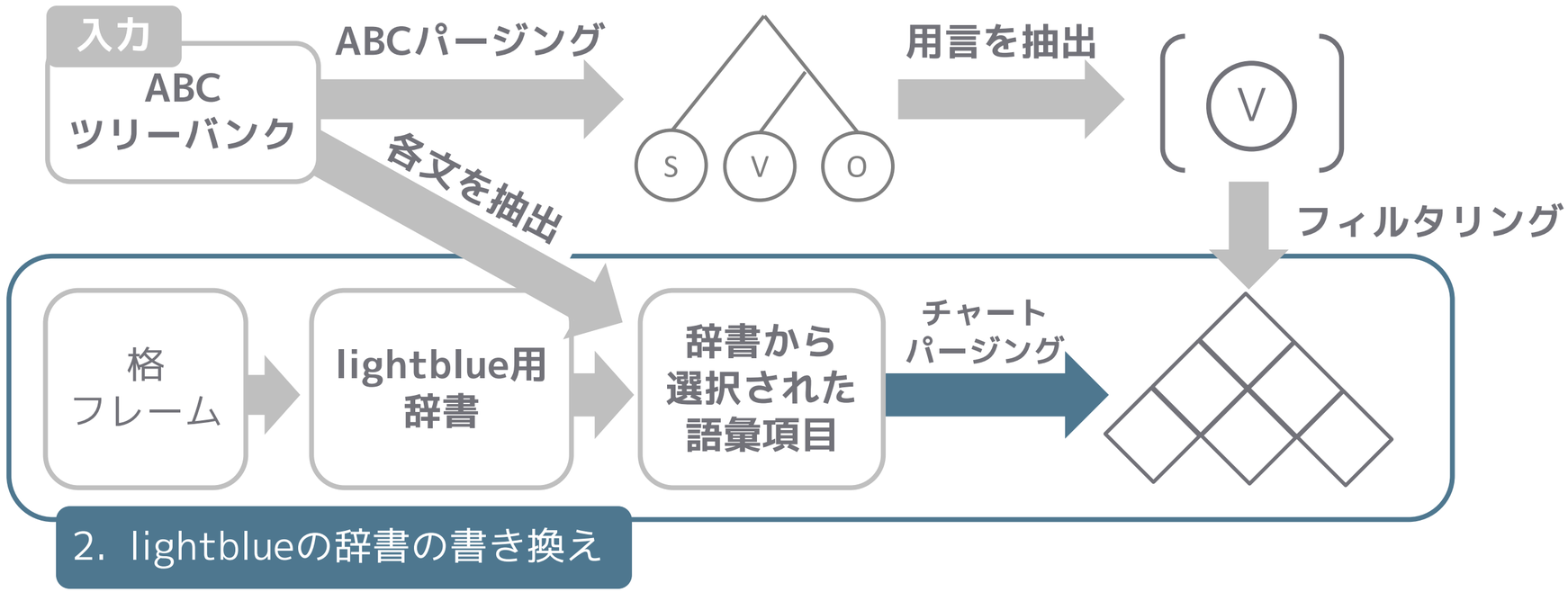
嫌いな人もいる

出力

嫌、い、な、人、も、い、る、
嫌い、いな、いる　などの**語彙項目**

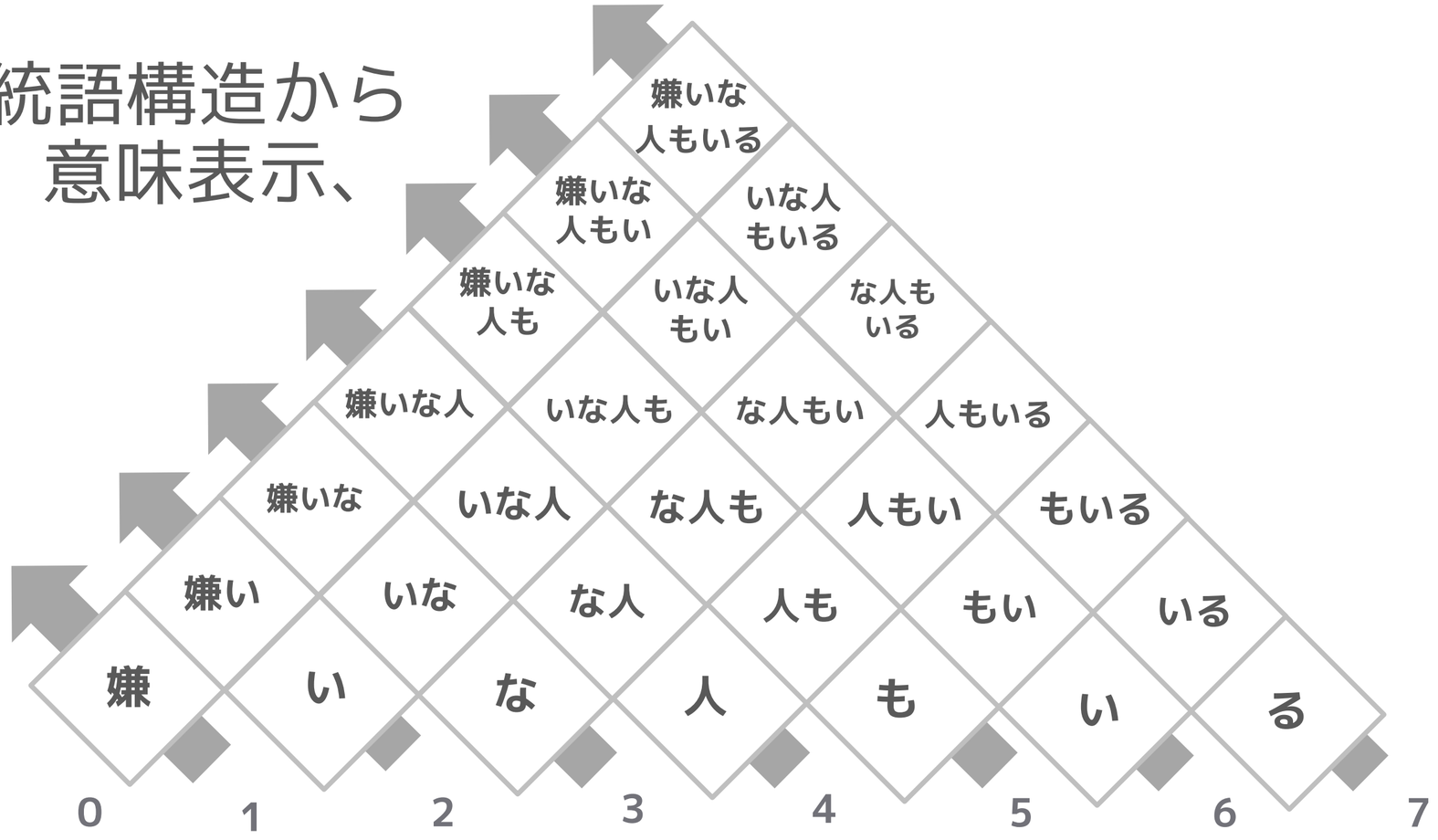
リフォージング – lightblueの辞書の書き換え

- 獲得した語彙項目を用いてチャートパージングを行う



left-corner chart parsing

- 部分統語構造を子の統語構造から作る際に、統語範疇、意味表示、スコアが計算される



left-corner chart parsing

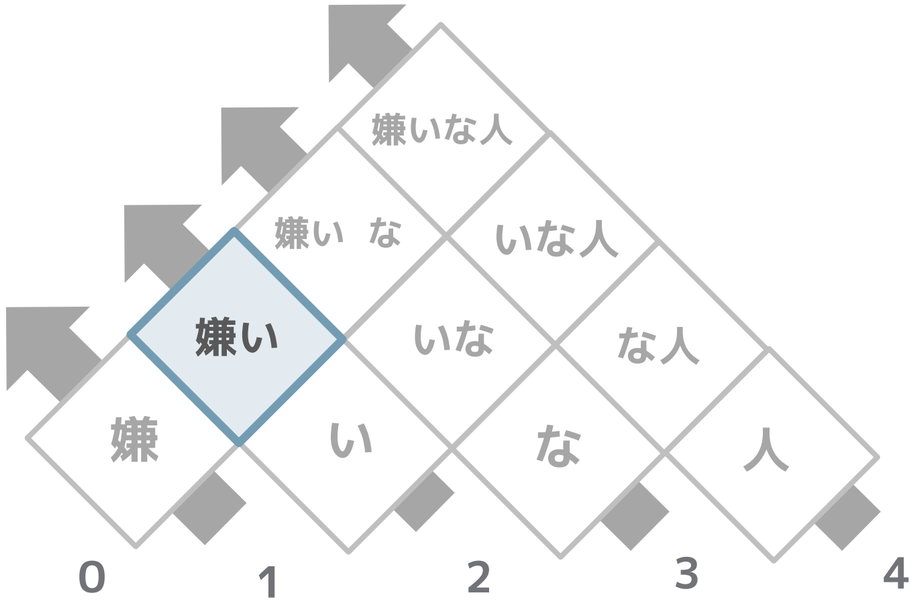
「嫌い」のマス(0,2)に登録されるのは、

1. **「嫌い」の統語情報 (状詞)**
2. 「嫌」＋「い」から合成できる統語情報（動詞の連用形）

例：(0,2)の語彙項目の一部

(“嫌い/きらい”,
$S_{[n:da|n:na|n:ni|+][nstem]} \setminus NP_{ga}$)

(“嫌う/きらう/ガオ”,
$S_{[v:5:w<1>][cont|mod:m]} \setminus NP_{ga} \setminus NP_{o}$)



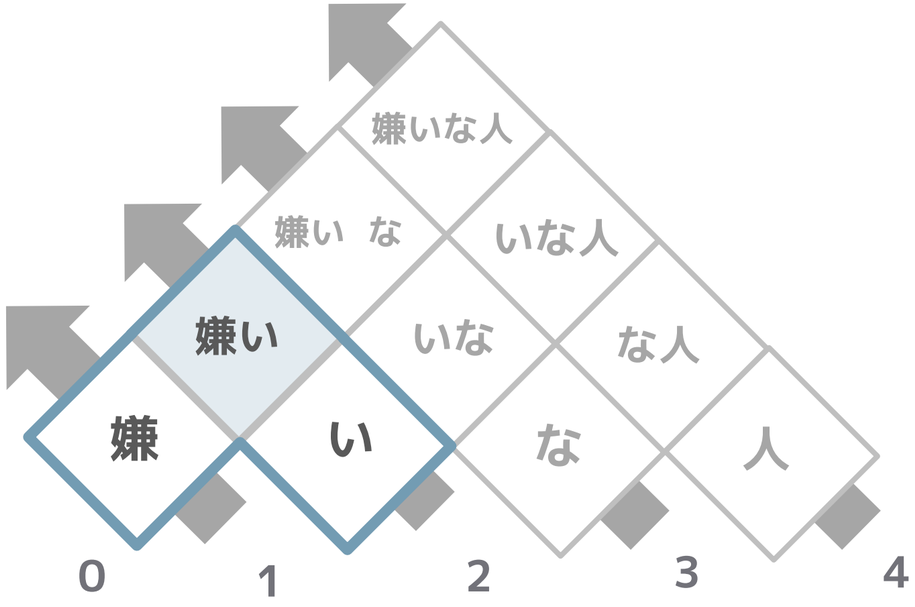
left-corner chart parsing

「嫌い」のマス(0,2)に登録されるのは、
1. 「嫌い」の統語情報 (状詞)
2. **「嫌」+「い」から合成できる統語情報（動詞の連用形）**

例：(0,2)の語彙項目の一部

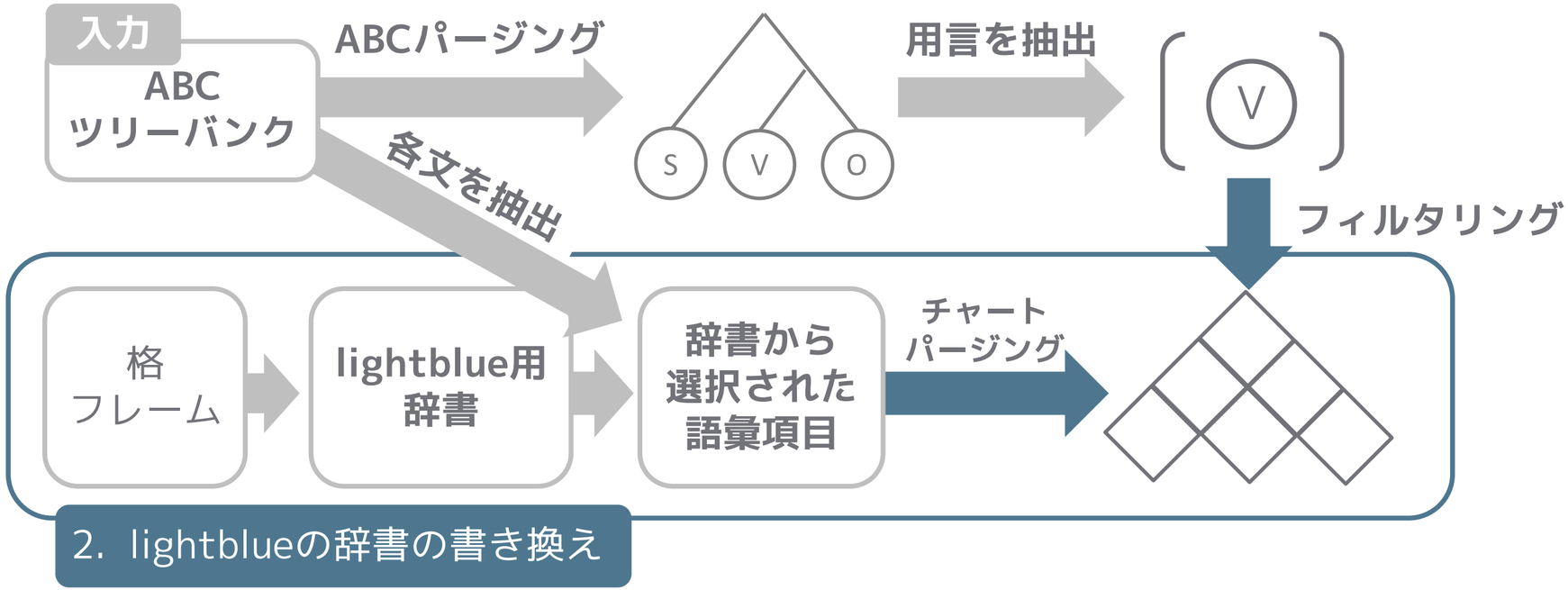
(“嫌い/きらい”,
$S_{[n:da|n:na|n:ni|+][nstem]} \setminus NP_{ga}$)

(“嫌う/きらう/ガオ”,
$\boldsymbol{S_{[v:5:w<1>][cont|mod:m]} \setminus NP_{ga} \setminus NP_{o}}$**)**

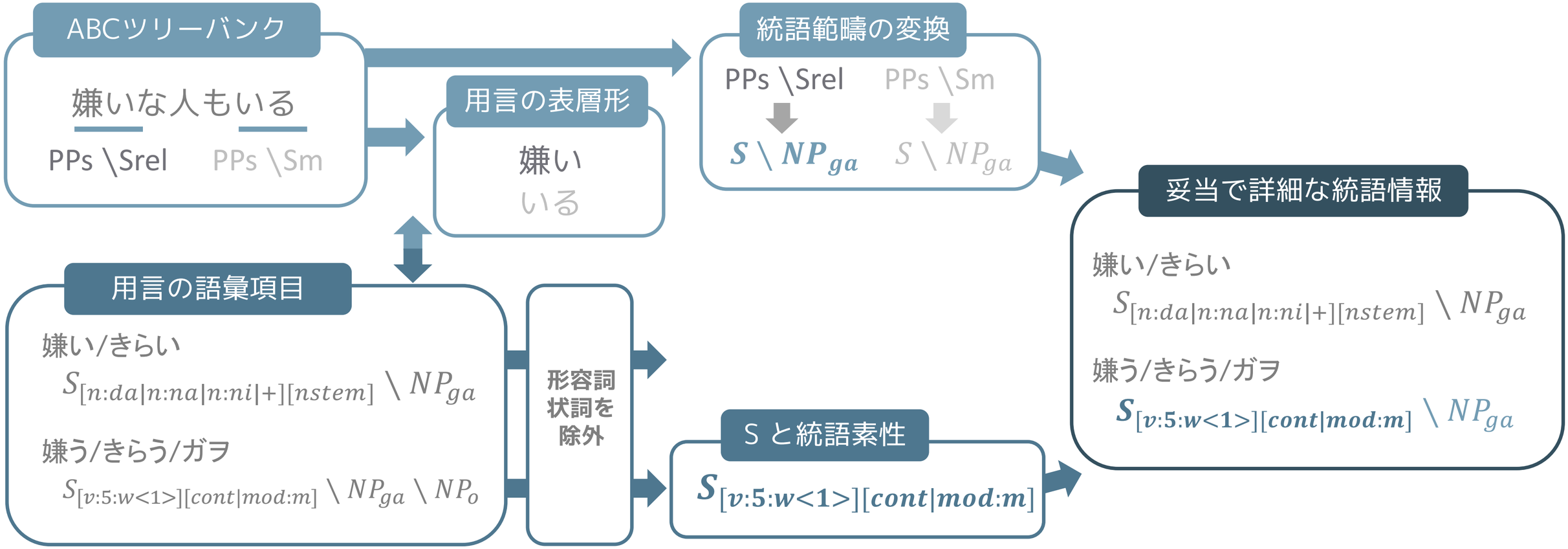


リフォージング – lightblueの辞書の書き換え

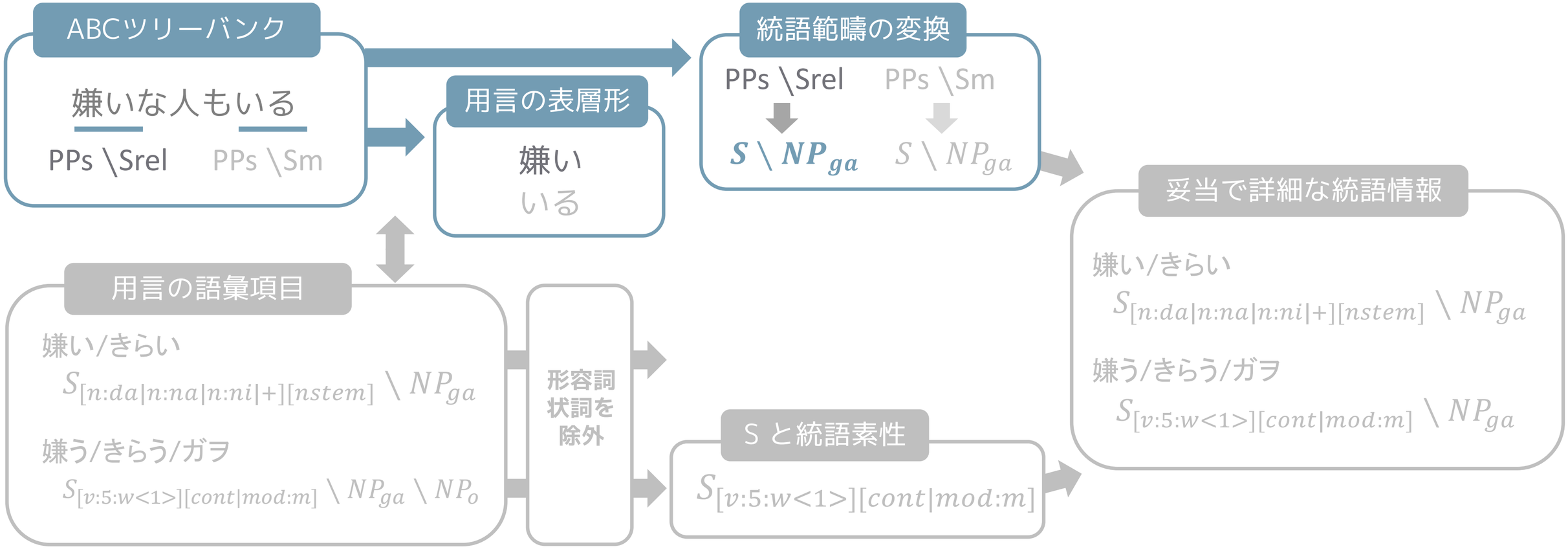
- 抽出した用言の項構造の情報でチャートの一部をフィルタリング



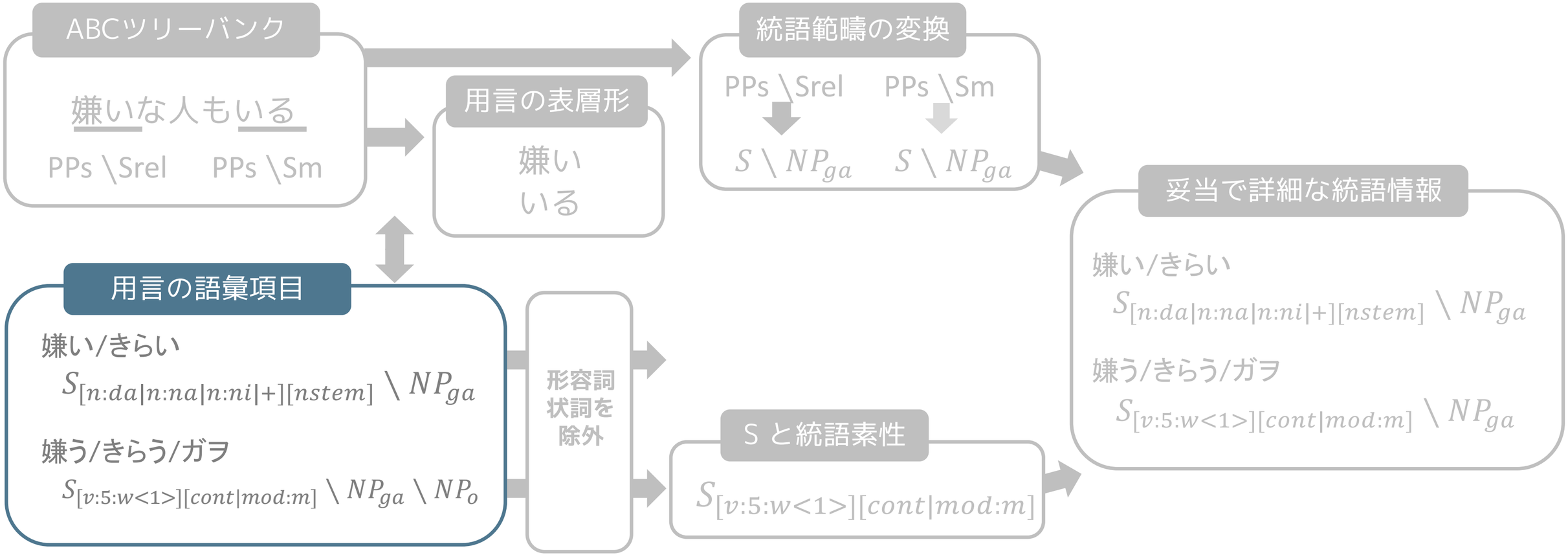
フィルタリングのアルゴリズム



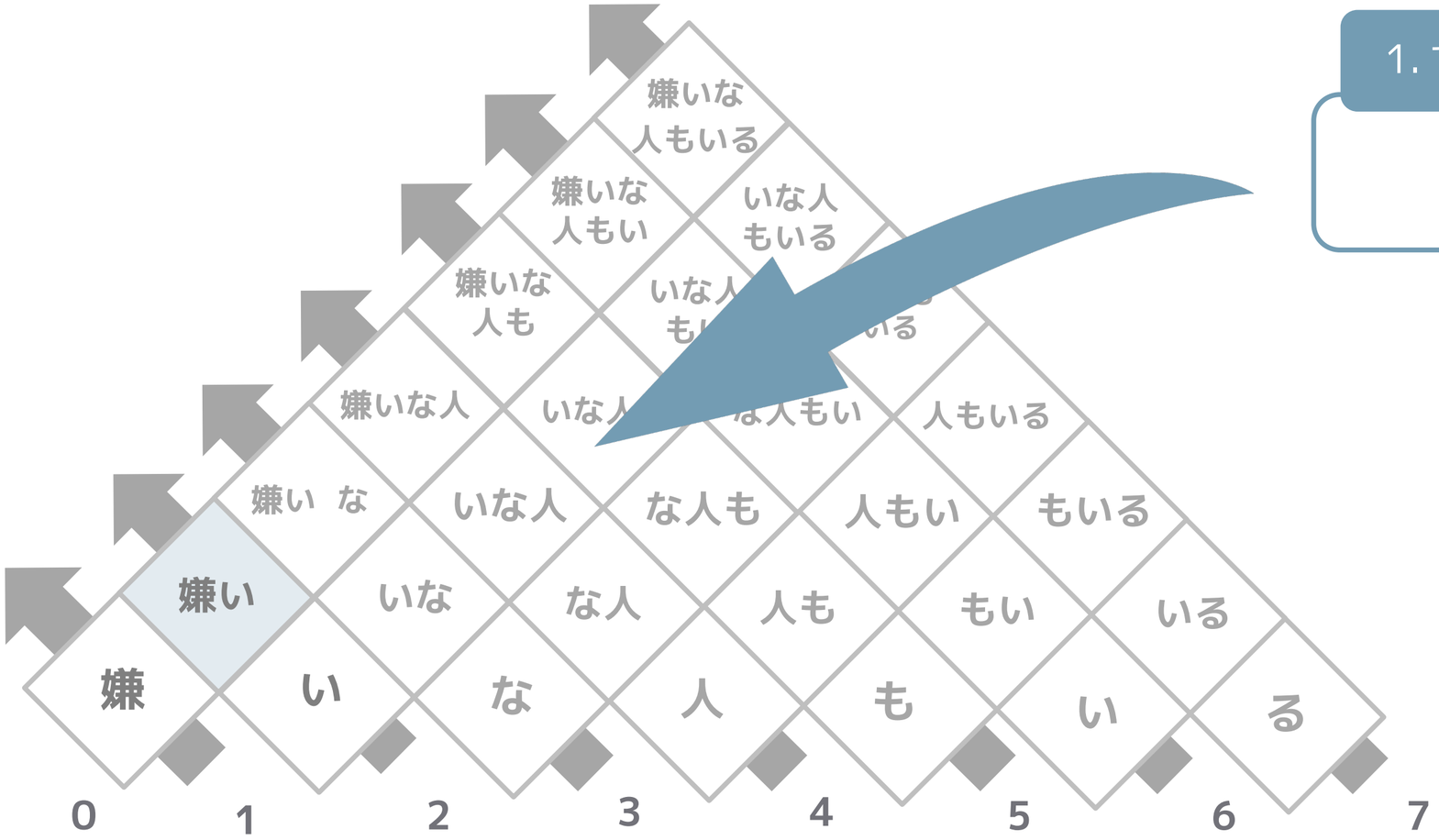
フィルタリングのアルゴリズム



フィルタリングのアルゴリズム



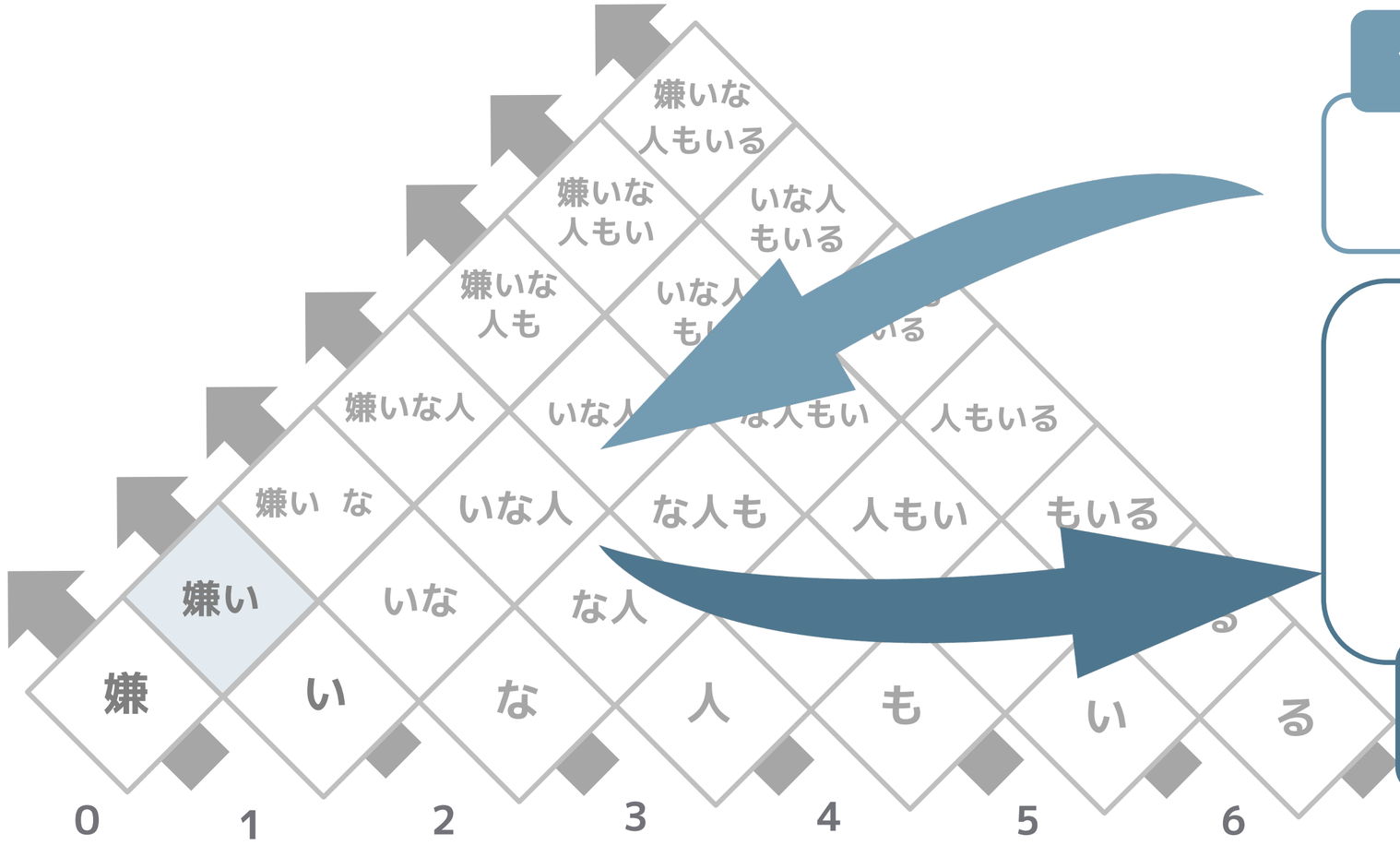
フィルタリングのアルゴリズム-語彙項目の取得



1. で抽出した用言の情報

[("嫌い", PPs \ SreI ,0,1)]

フィルタリングのアルゴリズム-語彙項目の取得



1. で抽出した用言の情報

["嫌い", PPs \ SreI ,0,1)]

嫌い/きらい

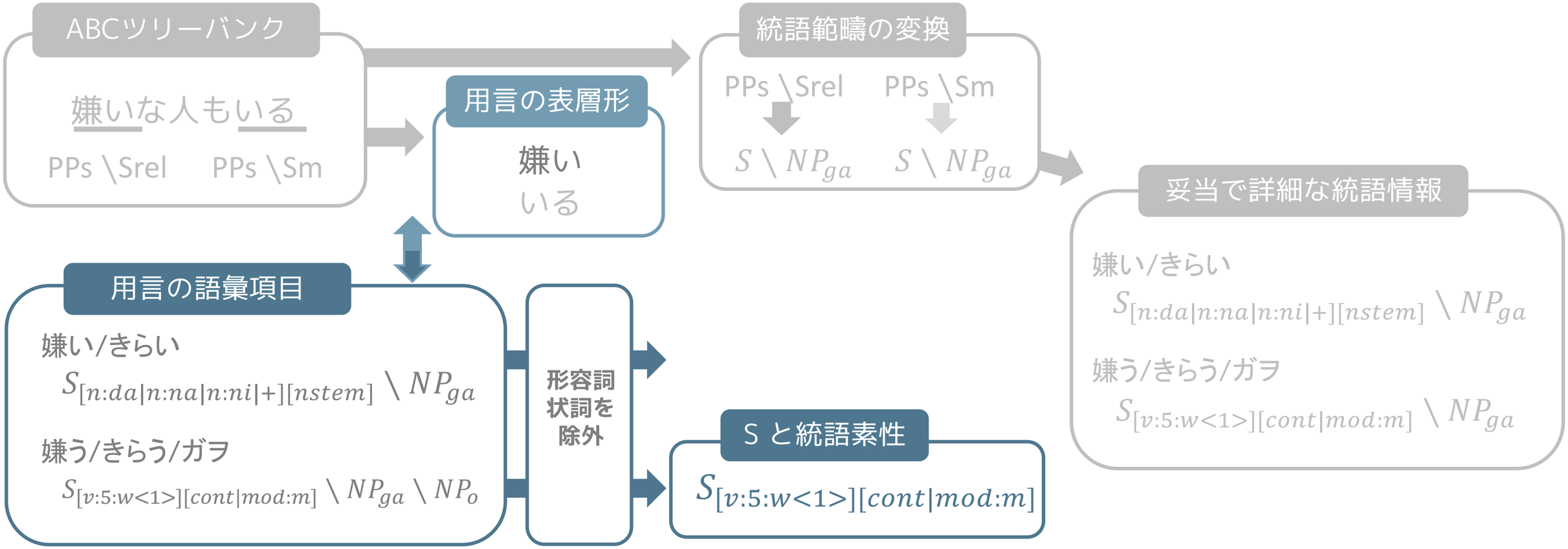
$S_{[n:da|n:na|n:ni|+][nstem]} \setminus NP_{ga}$

嫌う/きらう/ガヲ

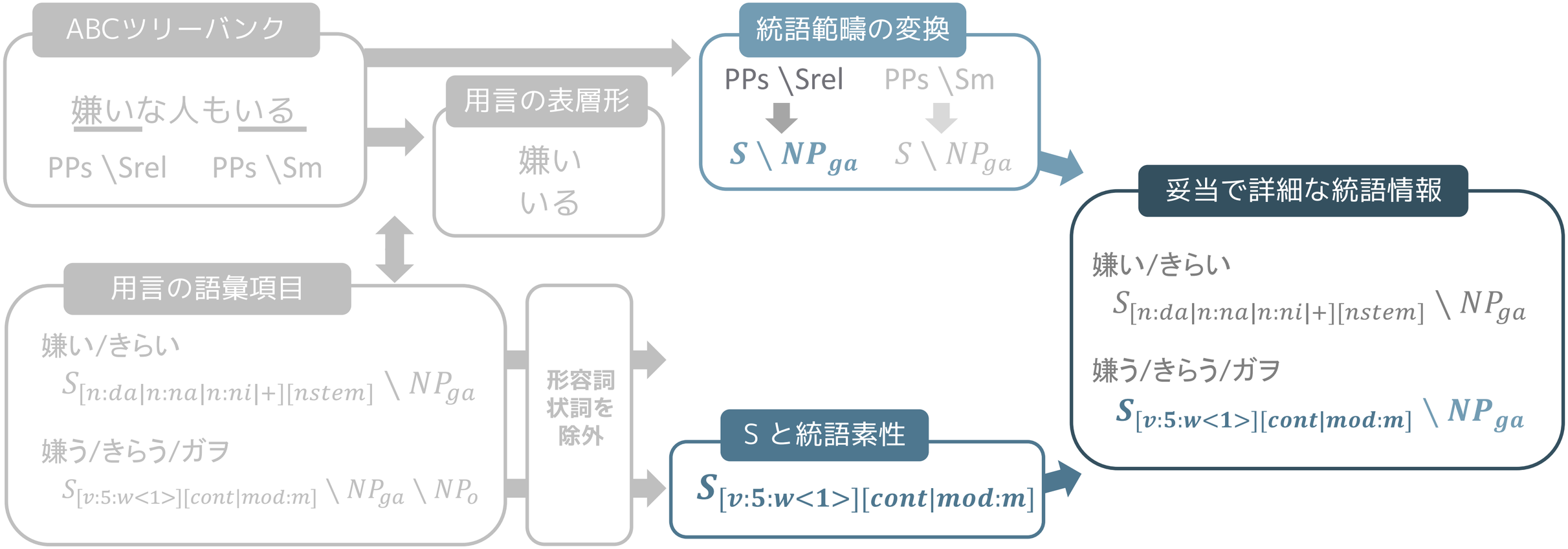
$S_{[v:5:w<1>][cont|mod:m]} \setminus NP_{ga} \setminus NP_{o}$

lightblueの辞書に登録されている
語彙項目の統語範疇

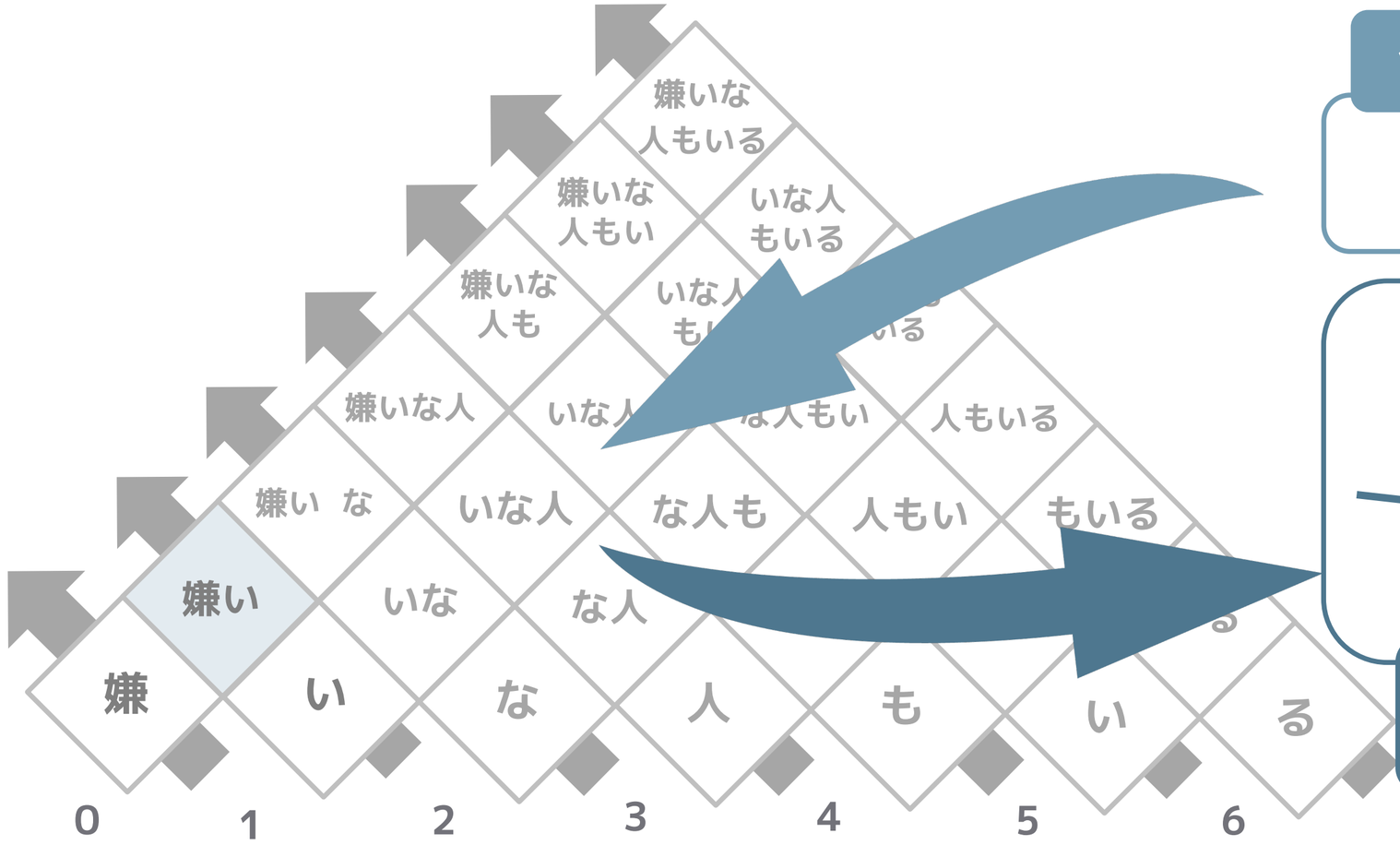
フィルタリングのアルゴリズム



フィルタリングのアルゴリズム



リフォージング – lightblueの辞書の書き換え



1. で抽出した用言の情報

[("嫌い", PPs \ SreI ,0,1)]

嫌い/きらい

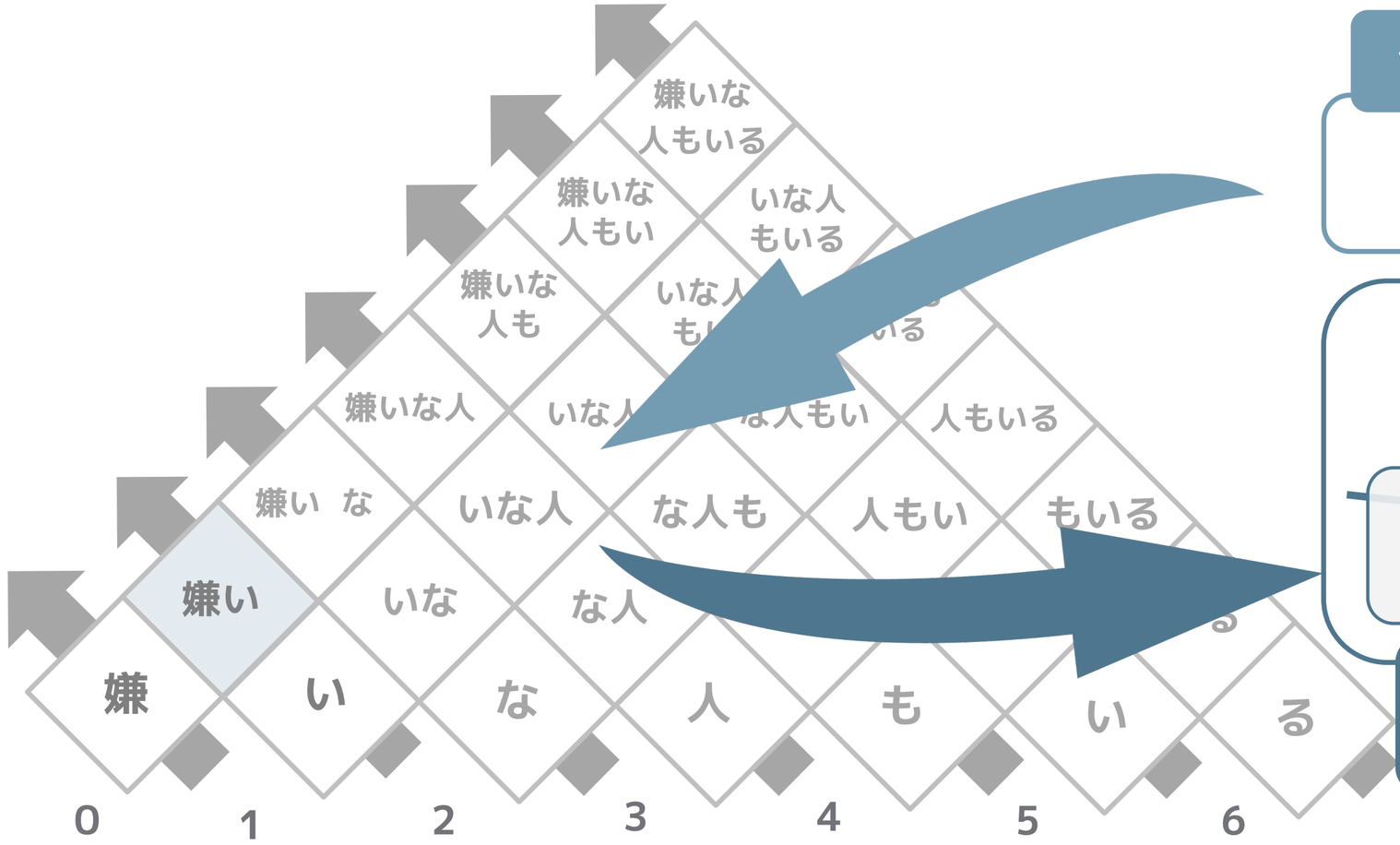
$S_{[n:da|n:na|n:ni|+][nstem]} \setminus NP_{ga}$

~~嫌う/きらう/ガヲ~~

~~$S_{[v:5:w<1>][cont|mod:m]} \setminus NP_{ga} \setminus NP_{o}$~~

lightblueの辞書に登録されている
語彙項目の統語範疇

リフォージング – lightblueの辞書の書き換え



1. で抽出した用言の情報

[("嫌い", PPs \ SreI ,0,1)]

嫌い/きらい

$S_{[n:da|n:na|n:ni|+][nstem]} \setminus NP_{ga}$

嫌う/きらう/ガ

$S_{[v:5:w<1>][cont|mod:m]} \setminus NP_{ga}$

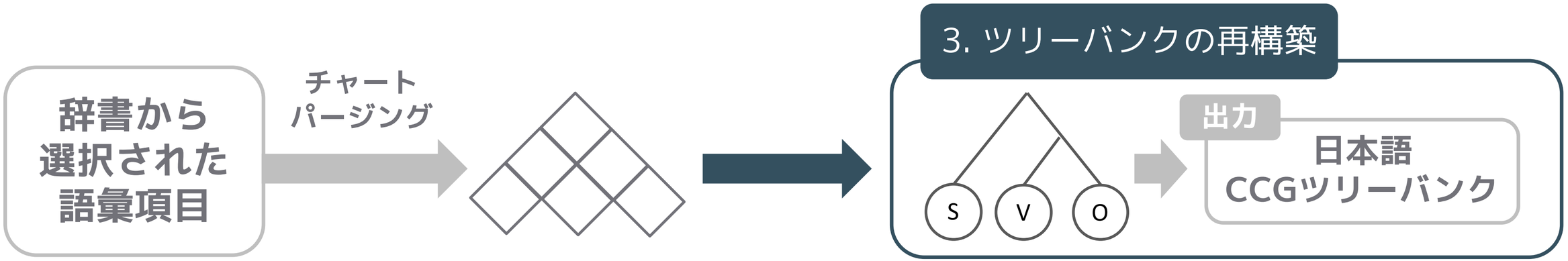
lightblueの辞書に登録されている
語彙項目の統語範疇

リフォージング – ツリーバンクの再構築

1. ABCツリーバンクからの用言抽出
2. lightblueの辞書の書き換え
3. ツリーバンクの再構築

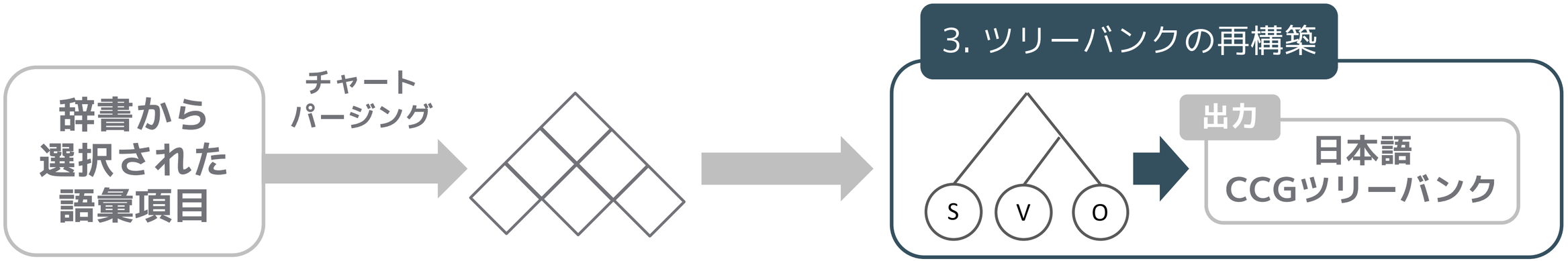
リフォージング – ツリーバンクの再構築

- フィルタリングされたチャートを用いてABCツリーバンクの文の統語解析を行う
- 解析が成功：CCG統語構造が出力



リフォージング – ツリーバンクの再構築

- ABCツリーバンクの全ての文で統語解析を行い、
  ツリーバンク形式にまとめる



リフォージング成功例

入力：会議が始まった

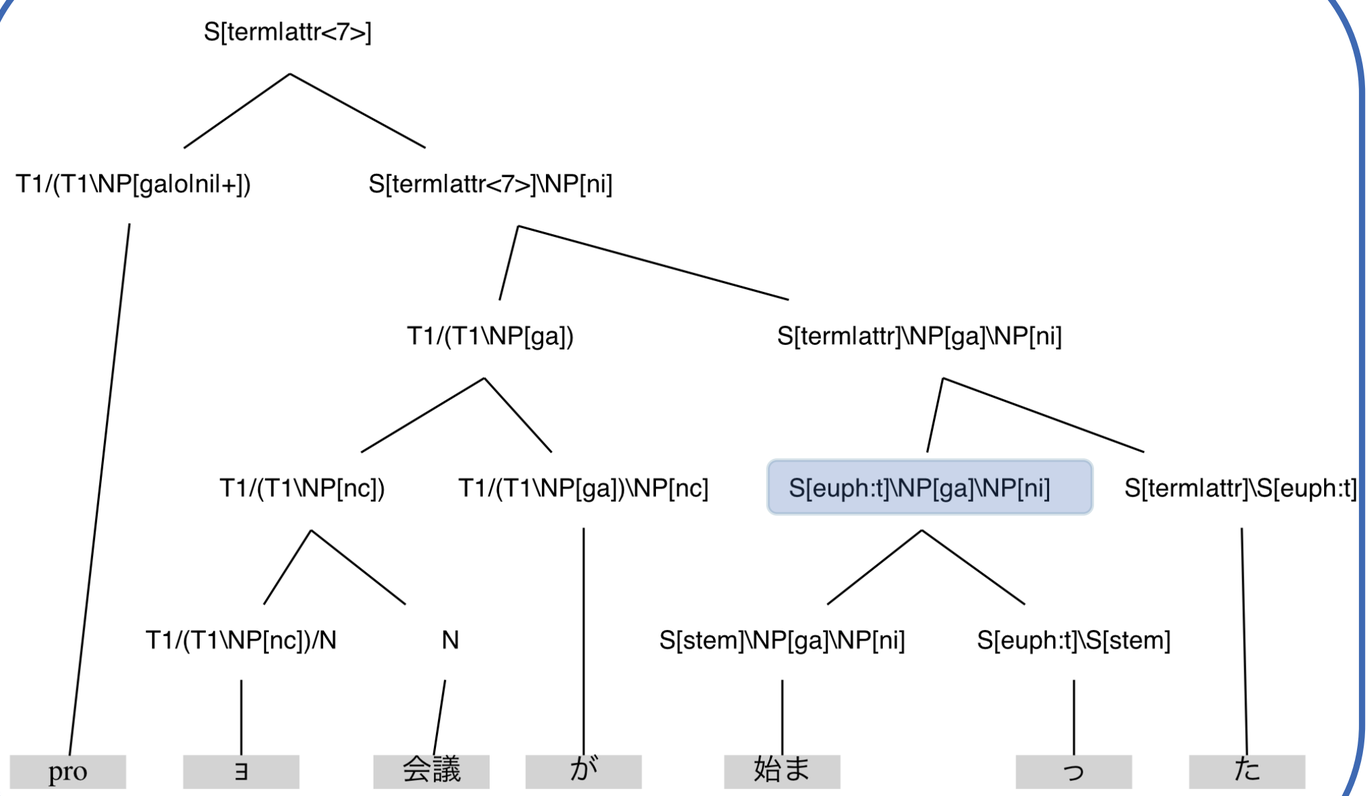
- lightblueがガ格と二格を必須格として分析している

lightblueの辞書に登録されている語彙項目の統語範疇

$$S_{[v:5:r<1>][euph:t]} \setminus NP_{ga} \setminus NP_{ni}$$

$$S_{[v:5:r<1>][euph:t]} \setminus NP_{ga}$$

フィルタリング前



リフォージング成功例

入力：会議が始まった

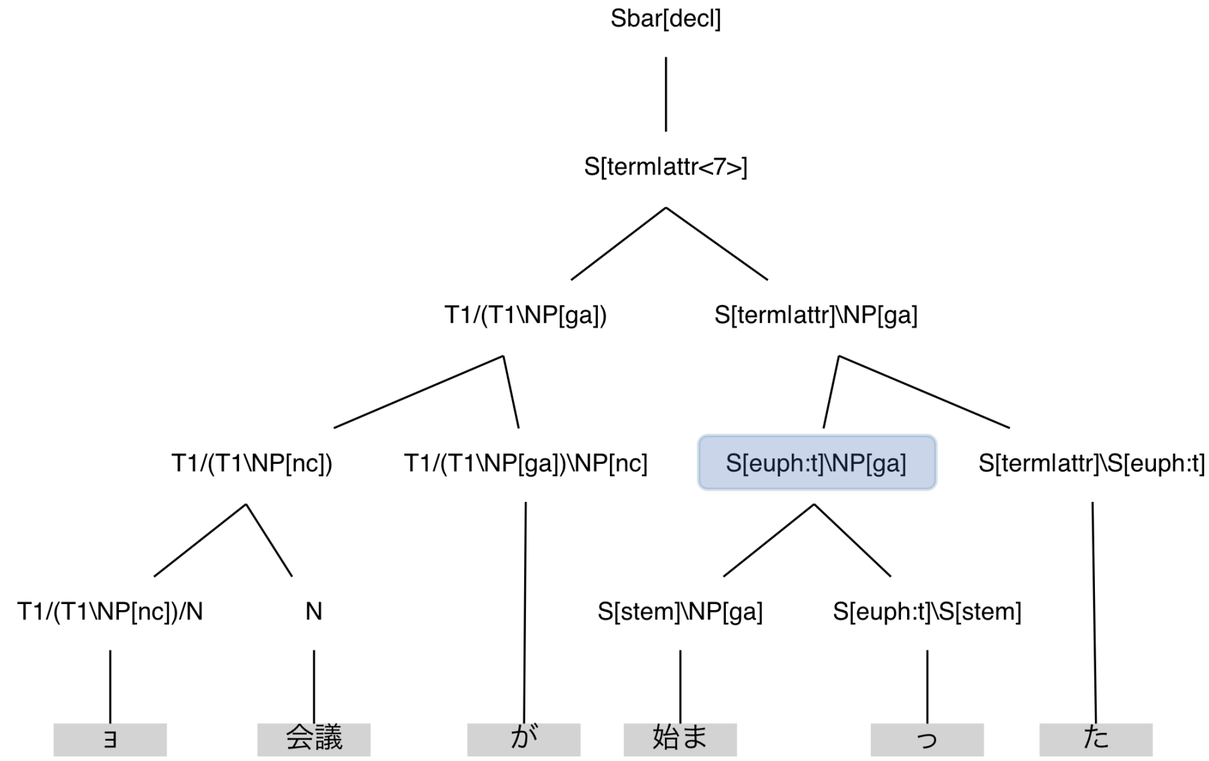
- 必須格がガ格のみに変更
- proが1つ削除された

lightblueの辞書に登録されている
語彙項目の統語範疇

$S_{[v:5:r<1>][euph:t]} \setminus NP_{ga}$

$S_{[v:5:r<1>][euph:t]} \setminus NP_{ga}$

フィルタリング後



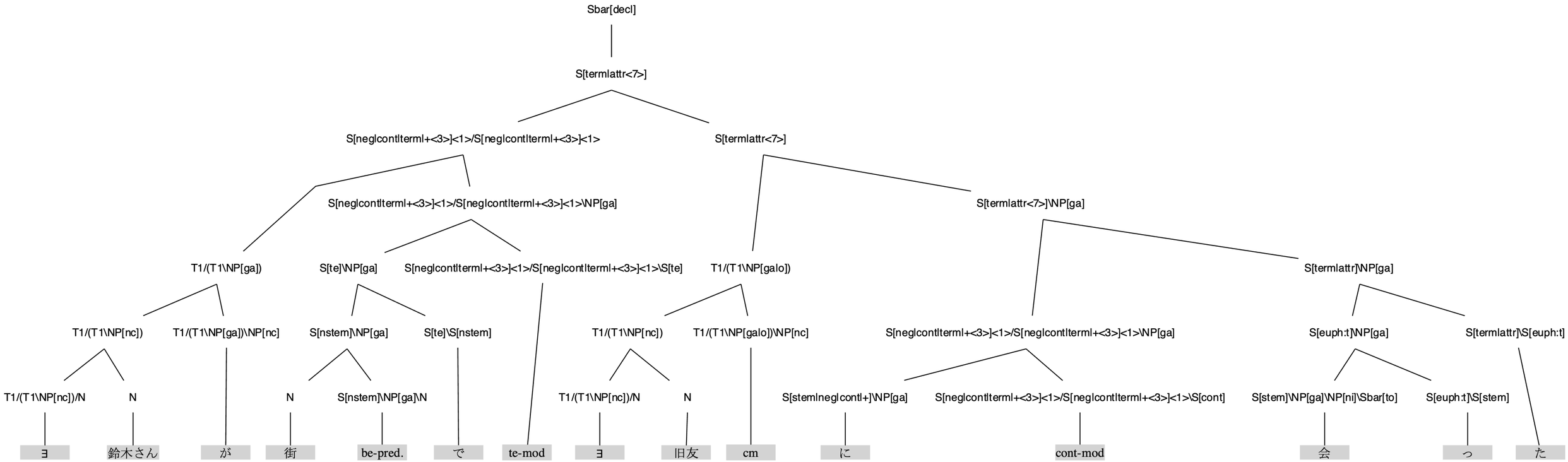
エラー分析

現在確認されているエラー

- ABCツリーバンクの項構造が誤っているケース
- lightblueの辞書のエントリが不足しているケース
- lightblueの外の関係・内の関係の分析が誤っているケース

エラー分析 – ABCツリーバンクの項構造が誤っているケース

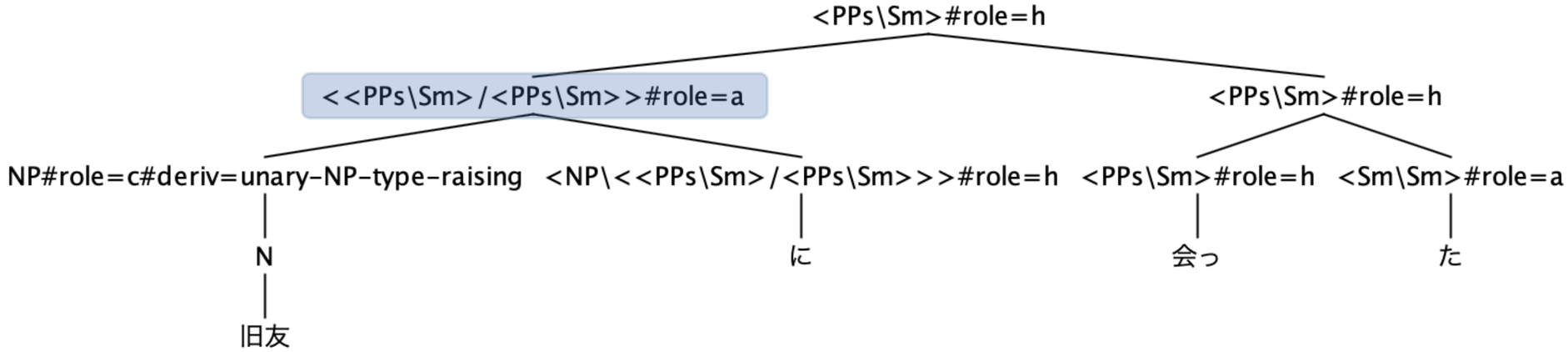
入力：鈴木さんが街で旧友に会った



エラー分析 – ABCツリーバンクの項構造が誤っているケース

入力：鈴木さんが街で旧友に会った

ABCツリーバンクの項構造

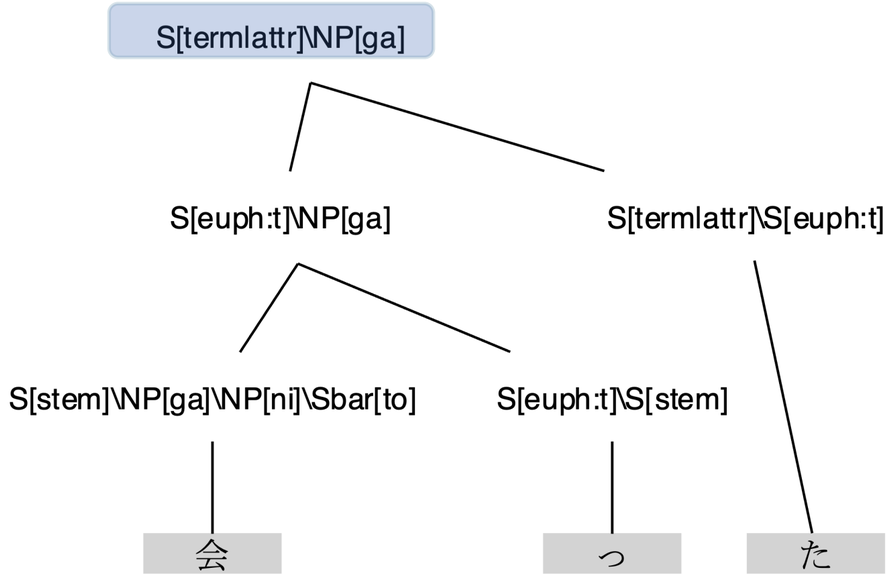


- 「旧友に」が**副詞句**として分析されている
- 正しくは**「ニ格」名詞句 PPo2** が付与されるべき

エラー分析 – ABCツリーバンクの項構造が誤っているケース

入力：鈴木さんが街で旧友に会った

lightblueの出力



「会っ」が**ガ格のみ**を取るように分析されてしまっている

正しくは**ガ格と二格**をとるべき

まとめ

- ABCツリーバンク と lightblue が持つ利点を組み合わせることで言語学的に妥当で詳細な統語情報を持った日本語CCGツリーバンクを構築する手法(リフォージング)を提案した。

- リフォージングによって正しい日本語CCG統語構造が部分的に得られたが、誤りが含まれるケースも残されている

今後の展望
- 正しい項構造をABCツリーバンク以外から取得する
- フィルタリングのアルゴリズムの改良をする

本研究は、
JST CRESTプロジェクト「**知識と推論に基づいて言語で説明できるAIシステム**」、
JSPS科研費 JP20K19868の支援を受けています

参考文献

1. Daisuke Bekki and Ai Kawazoe. Implementing variable vectors in a CCG parser.
   In Logical Aspects of Computational Linguistics. Celebrating 20 Years of LACL (1996–2016), pp. 52–67, Berlin, Heidelberg, 12 2016. Springer Berlin Heidelberg.
2. Daisuke Bekki and Hitomi Yanaka. Is Japanese CCGBank empirically correct? A case study of passive and causative constructions.
   In Proceedings of Treebanks and Linguistic Theories (TLT) 2023 (to appear), the workshop in the Georgetown University Round Table on Linguistics 2023 (GURT2023), forthcoming.
3. Daisuke Kawahara and Sadao Kurohashi. A fullylexicalized probabilistic model for Japanese syntactic and case structure analysis. In Proc. of the Human Language Technology Conference of the NAACL, Main Conference, June 2006.
4. Yoshikawa Masashi, Noji Hiroshi, and Matsumoto Yuji. A* CCG parsing with a supertag and dependency factored model.
   In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 277–287, Vancouver, Canada, 2017. Association for Computational Linguistics.
5. Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. Annotating a Japanese text corpus with predicate-argument and coreference relations. In Proceedings of Linguistic Annotation Workshop, pages 132–139.
6. Mark Steedman. The Syntactic Process. MIT Press, 2000.
7. Mark J. Steedman. Surface Structure and Interpretation. The MIT Press, Cambridge, 1996.
8. 植松すみれ, 松崎拓也, 花岡洋輝, 宮尾祐介, 美馬秀樹. 統語・意味コーパスの統合と再解釈による大規模な日本語CCG文法の開発.
   人工知能学会全国大会論文集, Vol. JSAI2013, pp. 4B11–4B11, 2013.
9. 戸次大介. 日本語文法の形式理論. くろしお出版, 東京, 2010.
10. 窪田悠介, 峯島宏次, 林則序, 岡野伸哉. 汎用的な範疇文法 ツリーバンクの構築.
    言語処理学会 第 25 回年次大会 発 表論文集 (2019 年 3 月), pp. 143–146. 一般社団法人 言 語処理学会, 2019.
11. 窪田悠介, 峯島宏次, 林則序, 岡野伸哉. ABC ツリーバン ク:学際的な言語研究のための基盤資源.
    言語処理学会 第 27 回年次大会 発表論文集 (2021 年 3 月), pp. 1529–1534. 一般社団法人 言語処理学会, 2021.
12. 花岡洋輝, 増田勝也, 植松すみれ, 美馬秀樹. 日本語助詞「と」コーパスの構築.
    言語処理学会 第 18 回年次大会 発 表論文集 (2012 年 3 月), pp. 247–250. 一般社団法人 言語処理学会, 2012.