3. グラフィカルインターフェースを使った検索式の作成

統語・意味解析コーパス (NPCMJ) チュートリアル@弘前大学

鈴木彩香・長崎郁*

2019.5.11

*本資料は、長崎郁「グラフィカルインターフェースを使った検索式の作成」（統語・意味解析コーパス (NPCMJ) チュートリアル@東北大学 2019.01.26）の改訂版である。

はじめに

- NPCMJ は、検索式を使って用例検索を行うことを前提としている
- 検索式
  - TGrep-lite
    /^*T*/ > ([^NP-SBJ] $ [^VB$])
  - XPath
    //node[matches(@word, '^*T*') and
    parent::node[matches(@cat, '^np-sbj') and
    ../node[matches(@cat, '^vb$')]]]
  - どちらも「**主語が関係節化されており、かつ関係節の述語が動詞である例**」を表す

はじめに

- TGrep-lite も XPath も使いこなすには記法をある程度知らなければならない
  - TGrep-lite の記法については次の時間で扱う
- 一方、NPCMJ には検索式の作成を補助するグラフィカルインターフェース(**Query builder**)も用意されている
- この時間は、具体的な例の検索を通じてQuery builder の使い方を体験する
  - **「の」でマークされた主語**
    「あなた**の** いる」
    「私**の** 知らない」
    「風**の** 吹く」...

検索までの手順

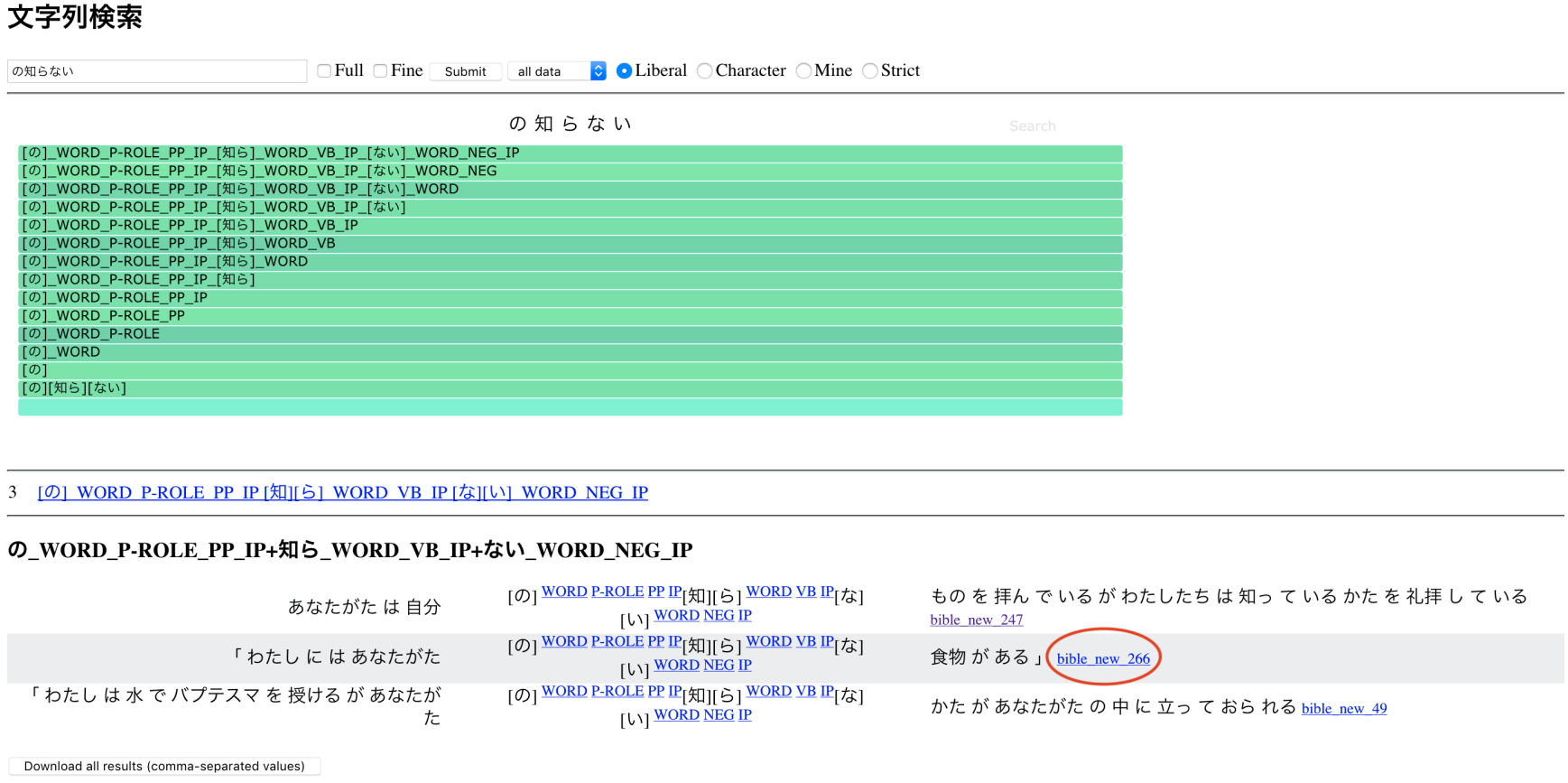
1. **実際の文例のツリーを表示させる**
   - **コーパスに存在する例**をもとに検索式を生成するため、事前に適切な例を見つけておく必要がある
     （以下では文字列検索を使う方法を紹介する）
2. **それを元に Query builder を立ち上げる**
3. **必要な情報を選び、検索式を自動生成させる**
4. **コーパスを検索**
5. **(必要に応じて検索式を修正し、再検索)**

1. 実際の文例のツリーを表示させる

- 「文字列検索」で「の+適当な動詞」の例を検索する
- 画面上部（ナビゲーションバー）の `文字列検索` をクリック、文字列検索画面へ

- ボックスに文字列を入力（今回は「の知らない」）、**Liberal** をチェック、`submit` ボタンをクリック

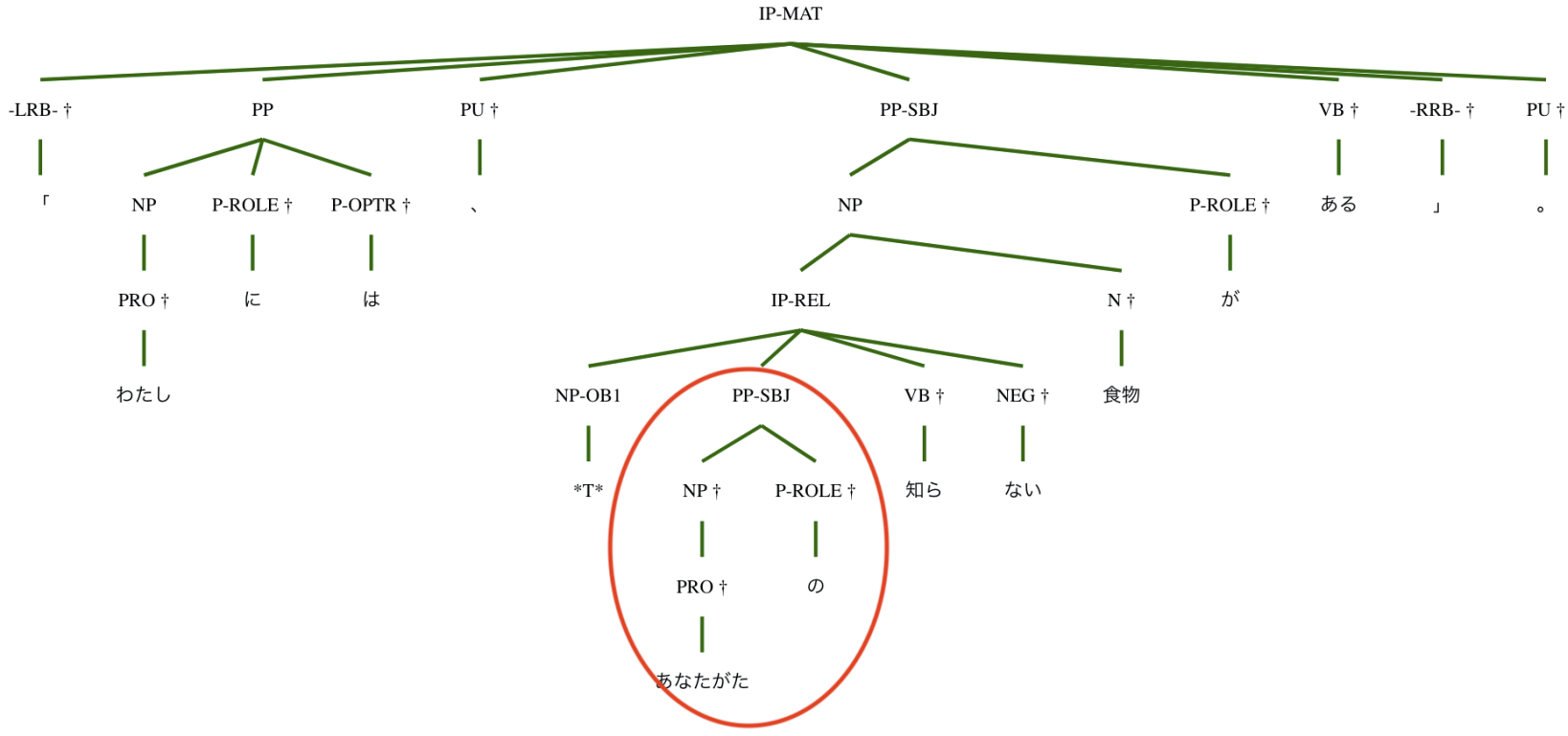
1.　実際の文例のツリーを表示させる



- ID をクリックするとツリーが表示される
  （今回はbible_new_266 をクリック）

bible_new_266

「 わたし に は 、 *T* あなたがた の 知ら ない 食物 が ある 」 。



- 助詞（P-ROLE）「の」の投射する助詞句（PP）に主語（SBJ）の拡張タグが付いていることに注目

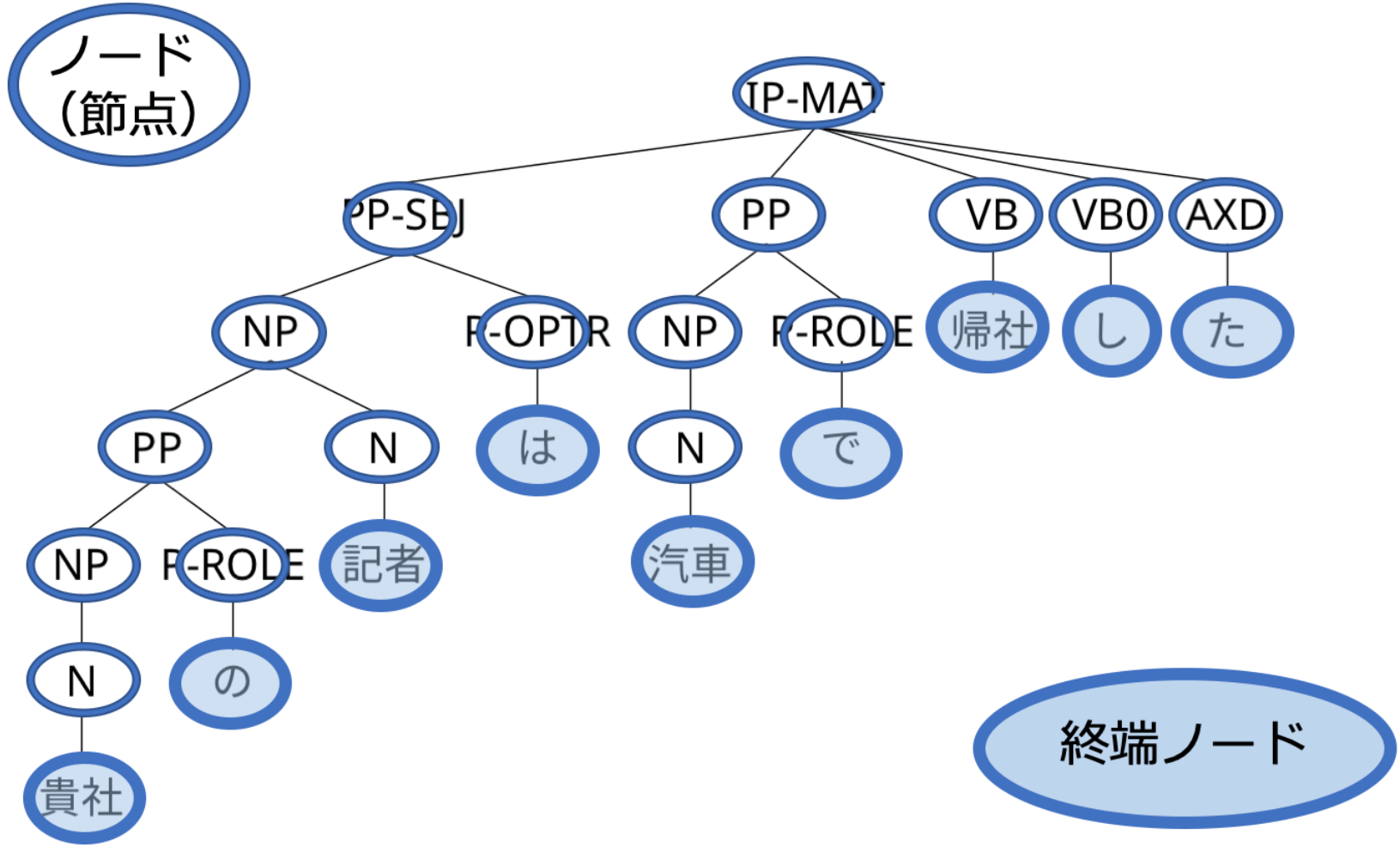
補足:「文字列検索」のオプション

	文字列の最初と最後はノードの境界と一致しなくてもよい	文字列の最初と最後はあるノードの先頭文字と末尾文字に一致する
文字列の中にノードの境界があるか否かを指定しない	Liberal	Character
文字列の中のノードの境界を半角スペースで指定する	Mine	Strict

補足:「文字列検索」のオプション

- “いろは” と入力した場合
  - **Character** – |い|ろ|は|、|いろ|は|、|い|ろは|、|いろは|
  - **Liberal** – 上記＋「い」と「は」が前後の別のノードの一部になっている例
  - **Strict** – |いろは| のみ
  - **Mine** – 上記に加えて、「いろは」がノードの一部をなしている例も検索

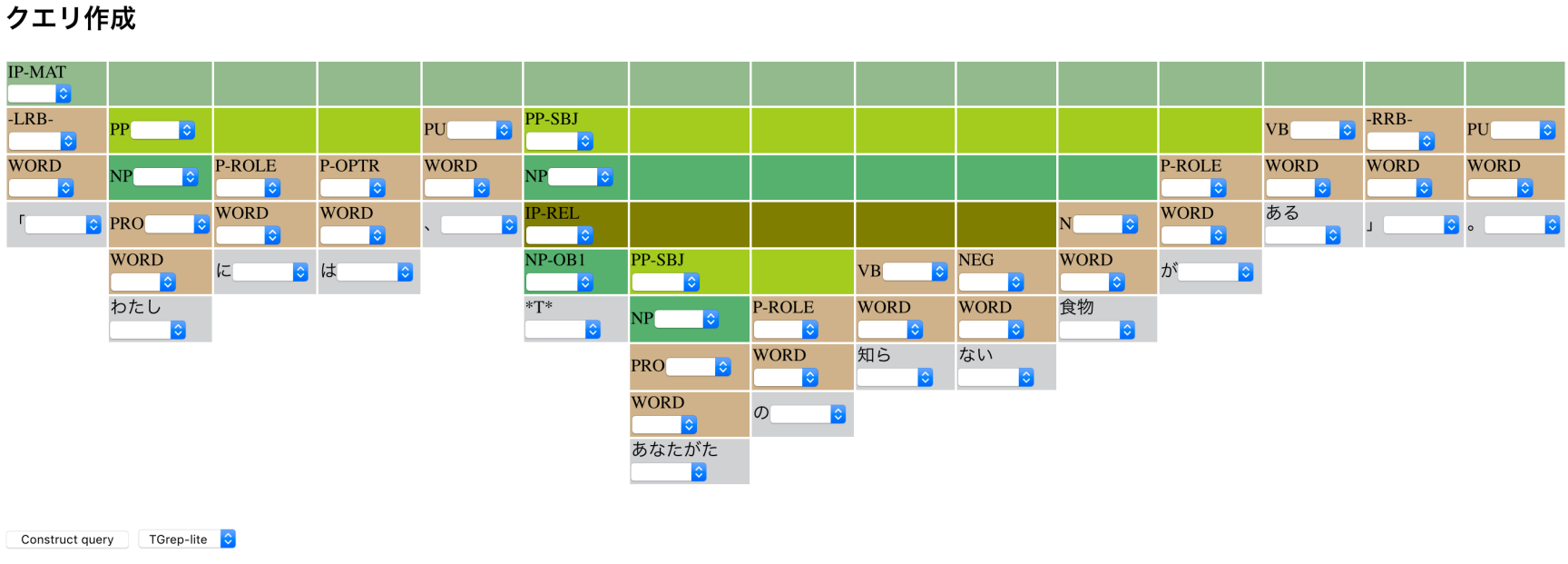
補足:「ノード(節点)」


ノード
(節点)
IP-MAT
PP-SBJ
PP
VB
VB0
AXD
NP
P-OPTR
NP
P-ROLE
帰社
し
た
PP
N
は
N
で
NP
P-ROLE
記者
汽車
N
の
貴社
終端ノード

2. Query builder を立ち上げる

- ツリーを表示させた状態で画面上部の `クエリ作成` をクリック

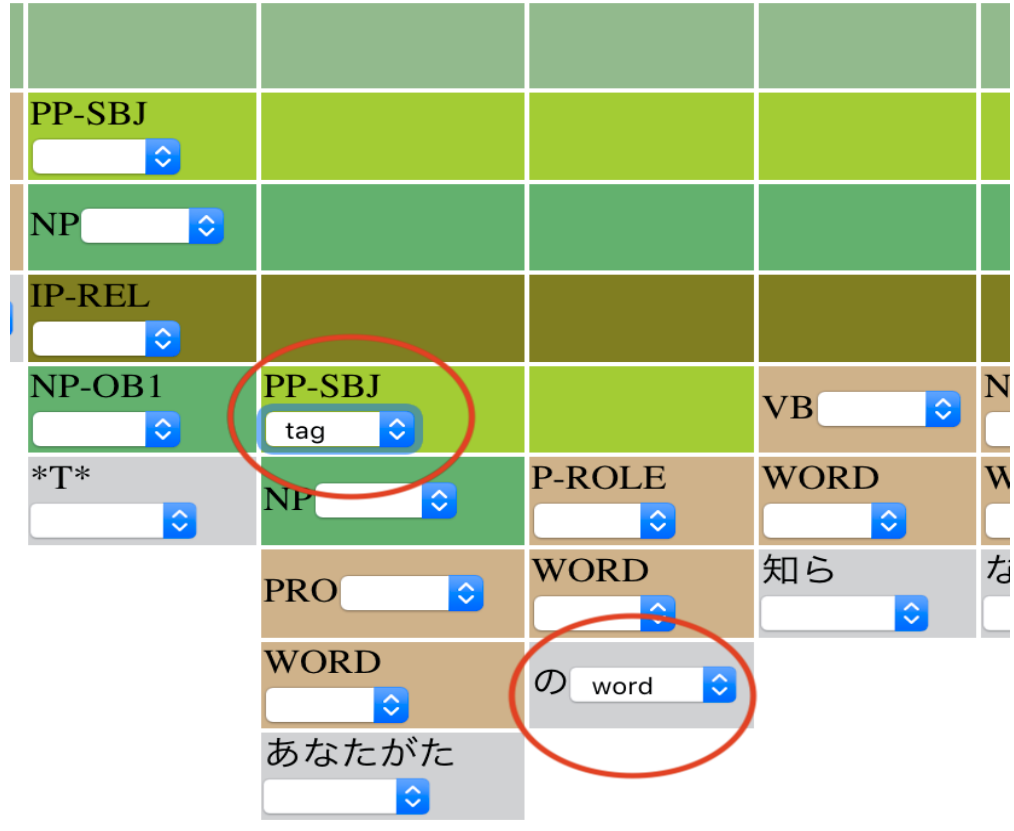
- Query builder が立ち上がる



3. 必要な情報を選択し、検索式を自動生成させる

1.「の」の横のドロップダウンリストをクリック、word を選ぶ

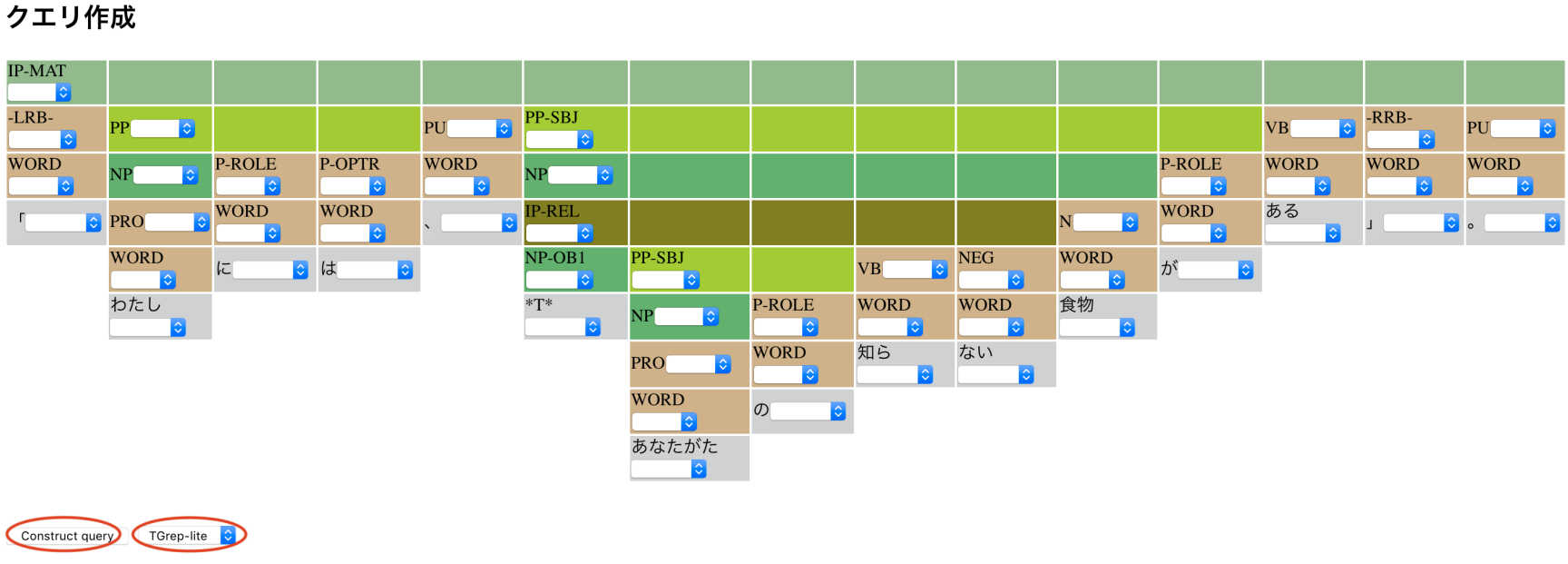
2. 左の列、PP-SBJ 横のドロップダウンリストをクリック、tag を選ぶ



3. 必要な情報を選択し、検索式を自動生成させる

3. 画面下 `Construct query` ボタンの横に TGrep-lite と表示されていることを確認

4. `Construct query` をクリック



3. 必要な情報を選択し、検索式を自動生成させる

- TGrep-lite の検索式がボックスに、ツリーにしたものがその下に表示される
- 検索式は「**PP-SBJ**（主語として機能する助詞句）で、それが **P-ROLE**（格助詞）を**直接支配**し、さらにその **P-ROLE** が「**の**」を**直接支配**する」という意味を表す

TGrep-lite クエリ（レビュー）

```
[PP-SBJ\b] < ([P-ROLE\b] < ({WORD} == /の\b/))
```

Submit | all data

4. コーパスを検索

- `submit` ボタンをクリックし、生成された検索式でコーパスを検索する
- 検索結果は、CSV format、Bracket format（括弧付きツリー）、Alpino XML format 形式でダウンロードできる

Query builder についてもう少し詳しく

- 終端ノードについては、ドロップダウンリストから 3 つのオプションを選ぶことができる
  - **word** – その文字列で終わる語形を指定
  - **!word** – その位置にあるのがその文字列で終わらない語形を指定

Query builder についてもう少し詳しく

- 非終端ノードの指定の仕方にも様々な選択肢がある
  - **tag** – そのタグ全体を指定
    （例：PP-SBJをtagとして指定
    「PP-OB1や単なるPPではなく**必ずPP-SBJであるもの**」）
  - **gen** – そのタグの基部を指定
    （例：PP-SBJをgenとして指定
    「PP-SBJだけでなく、PP-OB1やらPPやら**あらゆるPP**」）
  - **ext** – そのタグの拡張部を指定
    （例：PP-SBJをextとして指定
    「PP-SBJだけでなく、NP-SBJなどの**あらゆる-SBJ**」）
- さらに詳しいオプションの説明は、画面上部の `クエリ作成について` をクリック

5. 検索式の書き換え

- 生成された検索式は修正し、再利用することができる
  - **修正前：**

    [PP-SBJ\b] < ([P-ROLE\b] < ({WORD} == /の\b/))
  - **「に」でマークされた主語：**

    [PP-SBJ\b] < ([P-ROLE\b] < ({WORD} == /**に**\b/))
  - **「の」でない助詞でマークされた主語：**

    [PP-SBJ\b] < ([P-ROLE\b] < ({WORD} **!**== /の\b/))
  - **「の」でも「に」でもない助詞でマークされた主語：**

    [PP-SBJ\b] < ([P-ROLE\b] < ({WORD} **!**== /**(の|に)**\b/))

練習問題

1.「（映画）が好き」「（英語）ができる」のように、目的語がガでマークされている例を検索しなさい。

2.「（私の）知らない人」のように目的語が関係節化された例を検索しなさい。

3.「会いに（来た）」「買いに（行く）」のような、移動の目的を表す節を検索しなさい。