

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ぬ
 - 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

文字列検索

の知らない Full Fine Submit all data Liberal Character Mine Strict

1. 実際の文例のツリーを表示させる

文字列検索

の知らない Full Fine all data Liberal Character Mine Strict

の知らない

Search

```
[の]_WORD_P-ROLE_PP_IP_[知ら]_WORD_VB_IP_[ない]_WORD_NEG_IP
[の]_WORD_P-ROLE_PP_IP_[知ら]_WORD_VB_IP_[ない]_WORD_NEG
[の]_WORD_P-ROLE_PP_IP_[知ら]_WORD_VB_IP_[ない]_WORD
[の]_WORD_P-ROLE_PP_IP_[知ら]_WORD_VB_IP_[ない]
[の]_WORD_P-ROLE_PP_IP_[知ら]_WORD_VB_IP
[の]_WORD_P-ROLE_PP_IP_[知ら]_WORD_VB
[の]_WORD_P-ROLE_PP_IP_[知ら]_WORD
[の]_WORD_P-ROLE_PP_IP_[知ら]
[の]_WORD_P-ROLE_PP_IP
[の]_WORD_P-ROLE_PP
[の]_WORD_P-ROLE
[の]_WORD
[の]
[の][知ら][ない]
```

3 [\[の\]_WORD_P-ROLE_PP_IP_\[知ら\]_WORD_VB_IP_\[ない\]_WORD_NEG_IP](#)

の]_WORD_P-ROLE_PP_IP_[知ら]_WORD_VB_IP_[ない]_WORD_NEG_IP

あなたがたは自分

[の] [WORD_P-ROLE_PP_IP](#)[知][ら] [WORD_VB_IP](#)[な]
[い] [WORD_NEG_IP](#)

ものを捧んでいるがわたしたちは知っているかたを礼拝している
[bible_new_247](#)

「わたしにはあなたがた

[の] [WORD_P-ROLE_PP_IP](#)[知][ら] [WORD_VB_IP](#)[な]
[い] [WORD_NEG_IP](#)

食物がある」 [bible_new_266](#)

「わたしは水でバプテスマを授けるがあなたがた

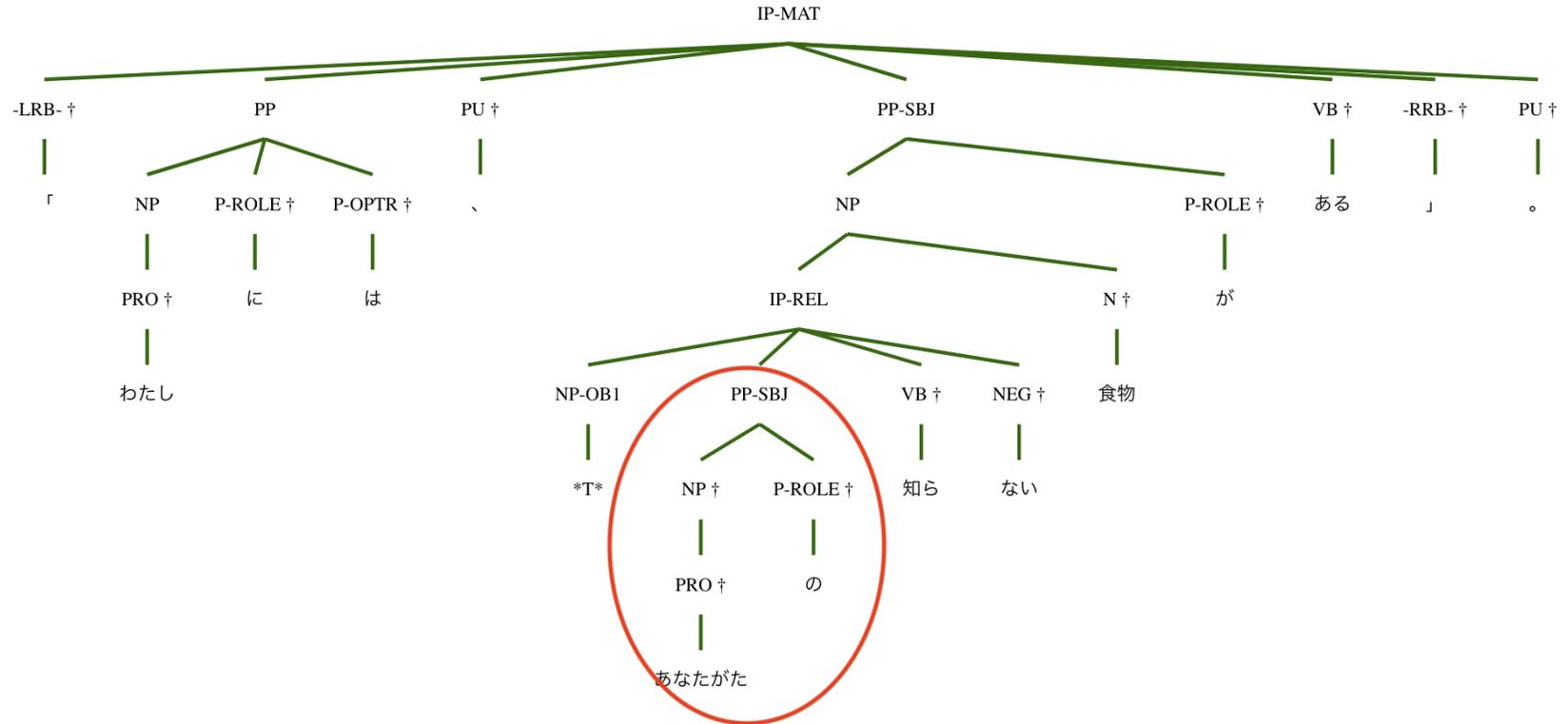
[の] [WORD_P-ROLE_PP_IP](#)[知][ら] [WORD_VB_IP](#)[な]
[い] [WORD_NEG_IP](#)

かたがあなたがたの中に立っておられる [bible_new_49](#)

- 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)



「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

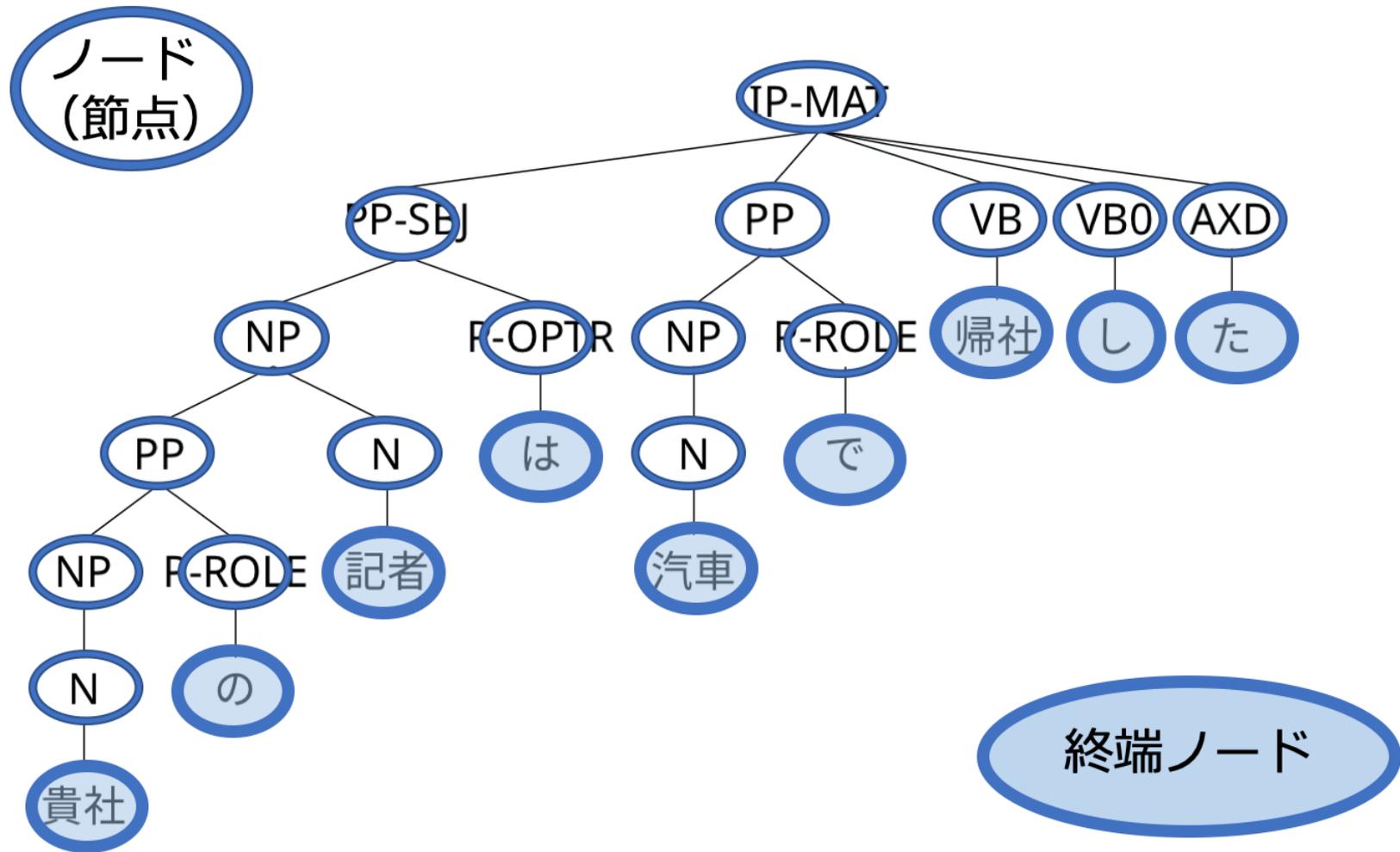
補足: 「文字列検索」のオプション

| | | |
|--------------------------|----------------------------|--------------------------------|
| | 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)」



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

| | | | | |
|--------|--------|--------|------|---|
| | | | | |
| PP-SBJ | | | | |
| NP | | | | |
| IP-REL | | | | |
| NP-OB1 | PP-SBJ | | VB | N |
| *T* | NP | P-ROLE | WORD | W |
| | PRO | WORD | 知ら | な |
| | WORD | | の | |
| | あなたがた | | | |

The image shows a grid-based search form interface. The grid has 5 columns and 10 rows. The first row is empty. The second row is labeled 'PP-SBJ' and has a dropdown menu. The third row is labeled 'NP' and has a dropdown menu. The fourth row is labeled 'IP-REL' and has a dropdown menu. The fifth row is labeled 'NP-OB1' and has a dropdown menu. The sixth row is labeled 'PP-SBJ' and has a dropdown menu with the value 'tag'. The seventh row is labeled '*T*' and has a dropdown menu. The eighth row is labeled 'NP' and has a dropdown menu. The ninth row is labeled 'P-ROLE' and has a dropdown menu. The tenth row is labeled 'WORD' and has a dropdown menu with the value 'の'. The eleventh row is labeled 'WORD' and has a dropdown menu with the value 'word'. The twelfth row is labeled 'あなたがた' and has a dropdown menu. Two red circles highlight the 'tag' dropdown in the sixth row and the 'word' dropdown in the tenth row.

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



の

PP-SBJ



P-ROLE



の

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。