統語・意味解析コーパス（NPCMJ）チュートリアル

セッション５

Tregex 検索式２

長崎 郁

2022.3.4

セッション５の内容

1. いろいろな関係結合子
2. 複雑な関係表現
3. 調べたい表現のアノテーションを知るには
4. その他
   Tregex と TGrep2 はどう違うの？
   NPCMJ の全データをダウンロードするには

セッション５の内容

1. いろいろな関係結合子
2. 複雑な関係表現
3. 調べたい表現のアノテーションを知るには
4. その他
   Tregex と TGrep2 はどう違うの？
   NPCMJ の全データをダウンロードするには

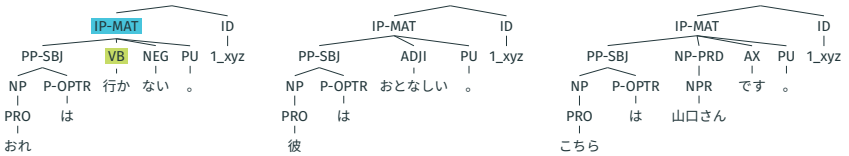
いろいろな関係結合子

復習

「A ＜ B」は「A が B を直接支配する」

```
/^IP\b/ < VB
```

IP が VB を直接支配する → 動詞を述語とする節



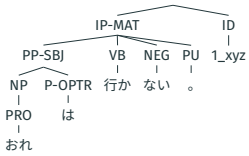
いろいろな関係結合子

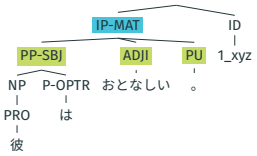
復習

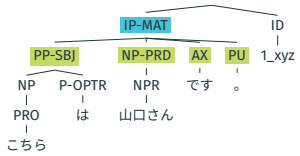
「A !< B」は「A が B を直接支配しない」

/^IP\b/ !< VB

IP が VB を直接支配しない → 動詞を述語としない節







いろいろな関係結合子

関係結合子には、ツリーにおける支配関係、前後関係、姉妹関係、イコール関係など、様々な関係を表すものがある。

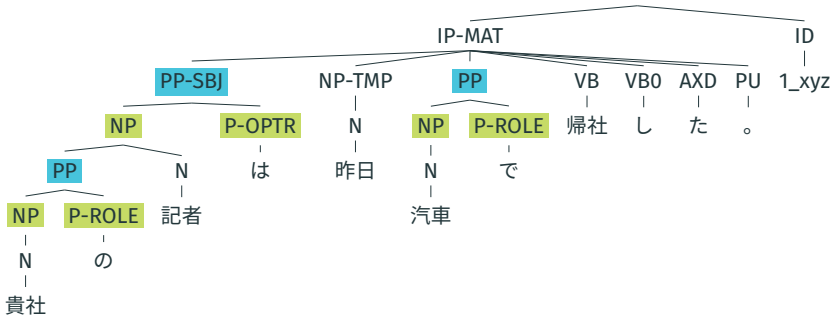
多くの関係結合子には順行・逆光の対があるため、マスターノードを柔軟に選ぶことができる。

いろいろな関係結合子 – 支配関係

A < B	A は B を直接支配する （A は B の親である）
A > B	A は B に直接支配される （A は B の子である）
A << B	A は B を支配する （A は B の先祖である）
A >> B	A は B に支配される （A は B の子孫である）

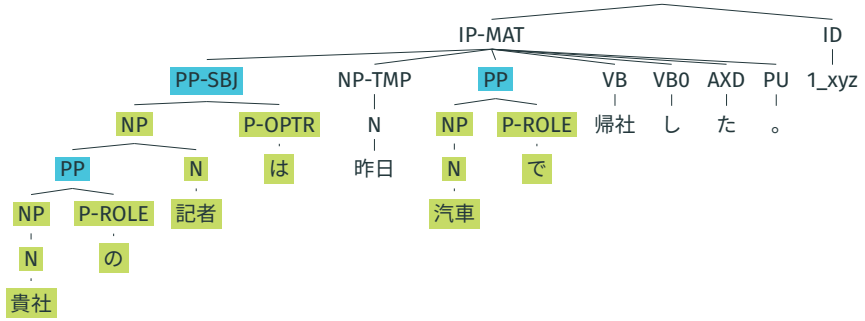
直接支配

/^PP\b/ < __（青がマスターノード）
__ > /^PP\b/（緑がマスターノード）



支配

/^PP\b/ << __（青がマスターノード）
__ >> /^PP\b/（緑がマスターノード）



練習１

以下の２つの検索式を使って、マスターノードが異なることを確認する。

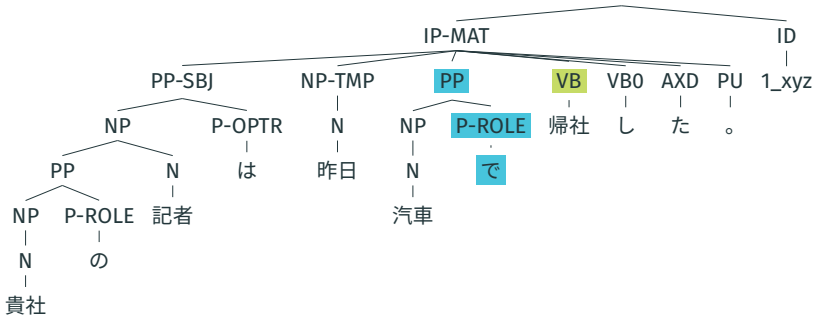
(1) VB < /^食べ/
(2) /^食べ/ > VB

いろいろな関係結合子 – 前後関係

A . B	AはBの直前に置かれる
A , B	AはBの直後に置かれる
A .. B	AはBに先行する
A ,, B	AはBに後続する

直前／直後

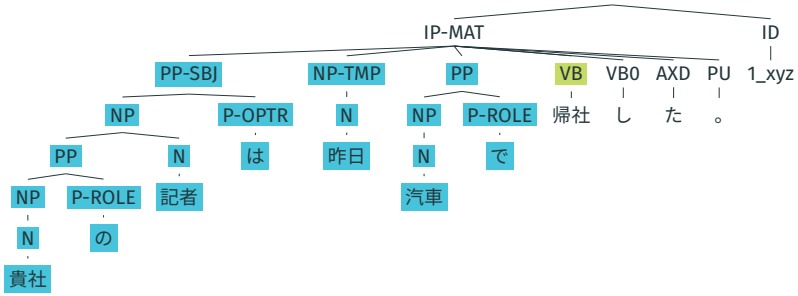
__ . VB（青がマスターノード）
VB , __（緑がマスターノード）



先行／後続

__ . . VB（青がマスターノード）

VB , , __（緑がマスターノード）



練習2

以下の2つの検索式を使って、マスターノードが異なることを確認する。

(1) て|で . /^(もら|貰)/

(2) /^(もら|貰)/ , て|で

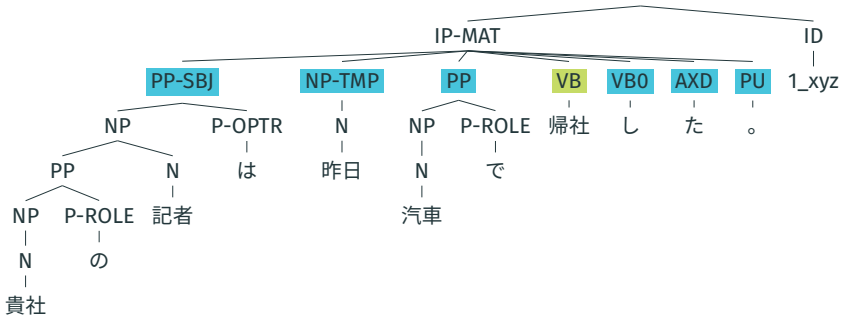
- どちらの検索式も動詞テ形に補助動詞「もらう」が続く例の抽出に使うことができる（ただし、「〜て・は・もらわない」のように間に助詞が現れる例は抽出されない）。

いろいろな関係結合子 – 姉妹関係

A $ B	AはBの姉妹である
A $. B	AはBの直前の姉妹である
A $, B	AはBの直後の姉妹である
A $.. B	AはBの先行する姉妹である
A $,, B	AはBに後続する姉妹である

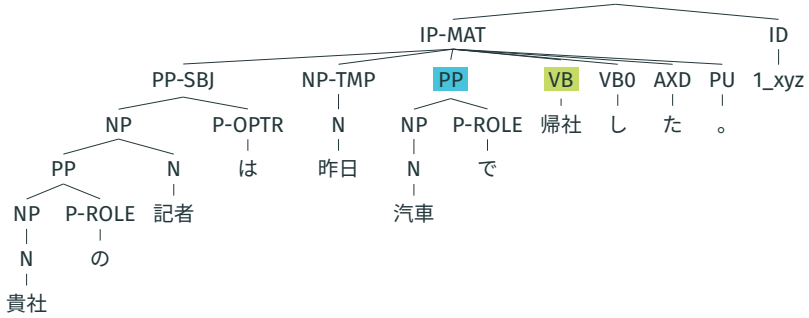
姉妹

__ $ VB（青がマスターノード）



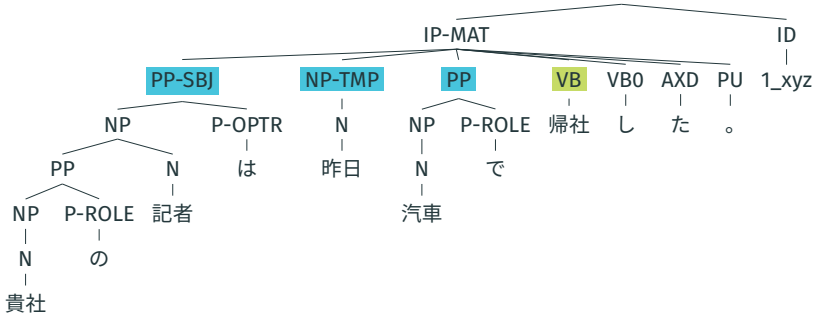
姉妹＋直前・直後

__ $. VB（青がマスターノード）
VB $, __（緑がマスターノード）



姉妹＋先行・後続

__ $.. VB（青がマスターノード）
VB $,, __（緑がマスターノード）



練習３

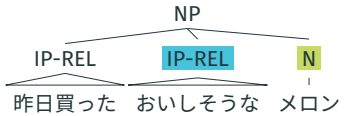
以下の２つの検索式はどのように異なるだろうか？

(1) `/^IP-REL\b/ $. /^N\b/`
(2) `/^IP-REL\b/ $.. /^N\b/`

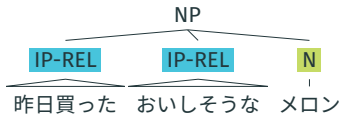
`/^IP-REL\b/` は関係節を、`/^N\b/` は普通名詞を表す。

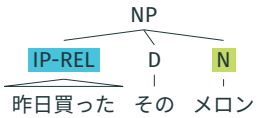
練習３（解答例）

`/^IP-REL\b/ $. /^N\b/`
→「関係節が普通名詞の直前の姉妹である」



`/^IP-REL\b/ $.. /^N\b/`
→「関係節が普通名詞の先行する姉妹である」





いろいろな関係結合子 – その他

A == B	A は B である
A <<, B	A は左端の子孫として B をもつ
A >>, B	A は B の左端の子孫である
A <<- B	A は右端の子孫として B をもつ
A >>- B	A は B の右端の子孫である
A <1 B (A <, B)	A は最初の子として B をもつ
A >1 B (B >, A)	A は B の最初の子である
A <-1 B (A <-, B)	A は最後の子として B をもつ
A >-1 B (B >-, A)	A は B の最後の子である
A <: B	A の唯一の子として B をもつ
A >: B	A は B の唯一の子である
A <<: B	A は B のみを支配する（枝分かれがない）
A >>: B	A は B のみによって支配される（枝分かれがない）

セッション5の内容

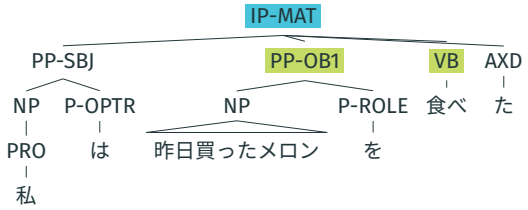
1. いろいろな関係結合子
2. **複雑な関係表現**
3. 調べたい表現のアノテーションを知るには
4. その他
   Tregex と TGrep2 はどう違うの？
   NPCMJ の全データをダウンロードするには

複雑な関係表現

ノード間の関係は2つ以上指定することもできる。

```
/^IP\b/ < /^PP-OB1\b/ < VB
```

IP（節）が PP-OB1 を直接支配し、かつ、VB を直接支配する



どちらの関係結合子も、マスターノードと、自身の後ろのノードとの関係を表すことに注意。

複雑な関係表現

```
/^IP\b/ < /^PP-OB1\b/ < VB
```

上の検索式は、以下のように書き換えが可能

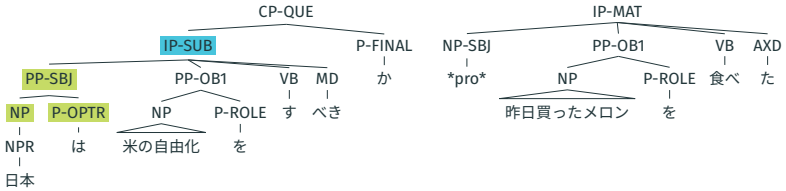
(a) `/^IP\b/ < /^PP-OB1\b/ & < VB`

(b) `( /^IP\b/ < /^PP-OB1\b/ ) < VB`

- 「かつ（&）」は省略できる。
- (. . .) で関係結合子を挟んだノードを囲って、グループ化できる。
  (. . .) はむしろ、次スライドのように使うことが多い。

複雑な関係表現

IP が *x*-SBJ を直接支配し、かつ、*x*-SBJ は空要素でない

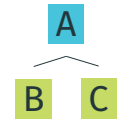



```
/^IP/ < (/-SBJ\b/ !< /^\*/)
```

括弧で囲むと、その中の関係結合子は自身の直前のノードと直後のノードとの関係を表すことになる。

複雑な関係表現

A < B < C



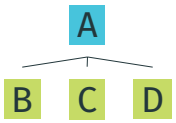
A < B かつ A < C

A < (B < C)

A < B かつ B < C

複雑な関係表現

A < B < C < D



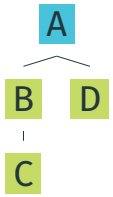
A < B かつ A < C かつ A < D

A < (B < (C < D))

A < B かつ B < C かつ C < D

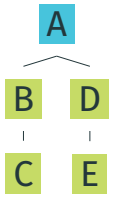
複雑な関係表現

A < (B < C) < D



A < B かつ B < C かつ A < D

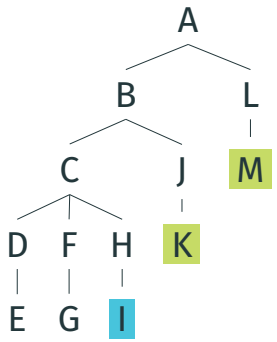
A < (B < C) < (D < E)



A < B かつ B < C
かつ A < D かつ D < E

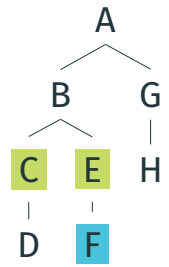
複雑な関係表現

```
I . (K . M)
```



I . K かつ K . M

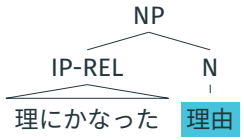
```
F > (E $, C)
```

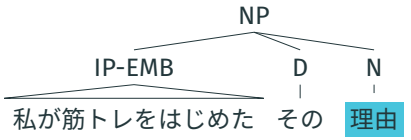


F > E かつ E $, C

練習 4

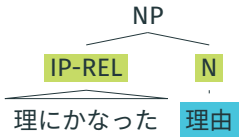
「理にかなった 理由」「私が筋トレをはじめた その 理由」のような、「理由」が名詞修飾節を伴った例を抽出したい。どのような検索式にしたらいいだろうか。





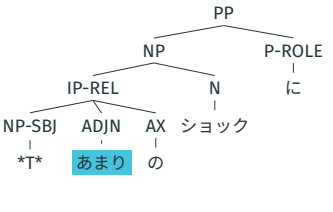
練習 4（解答例）

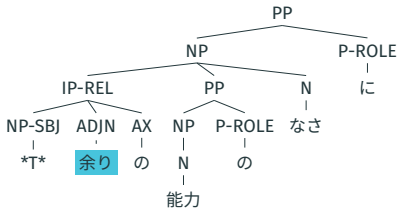
(a) 理由 > (N $,, /^(IP-REL|IP-EMB)\b/)
(b) 理由 > (N $,, /^IP\b/)



練習 5

「あまりのショックに」「余り能力のなさに」に共通する「あまり の 〜 に」を含む例を抽出したい。どのような検索式にしたらいいだろうか。



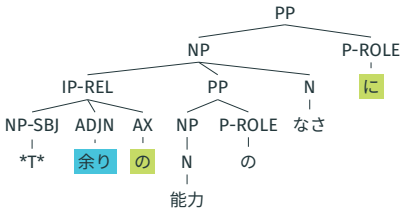


練習５（解答例）

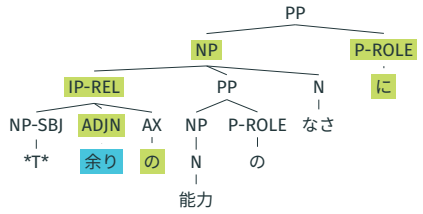
(a) あまり|余り . (の .. に)

(b) あまり|余り . の > (ADJN > (/^IP-REL\b/ > (/^NP\b/ $.. (P-ROLE < に))))

(a) の場合



(b) の場合



練習５（解答例）

(a) の検索式はシンプルで手軽だが、意図しない例も抽出される可能性がある。

> … 十 人 余り の 見物 が 一 かたまり に なっ て…

以下のように１つのツリーの中で、２回ヒットすることがある。そのうちの一方は意図しない例。

> あまり の ショック に 、彼女 は その 場 に 倒れ込ん だ
>
> あまり の ショック に 、彼女 は その 場 に 倒れ込ん だ

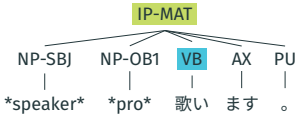
(b) の検索式では意図しない例は出てこない。

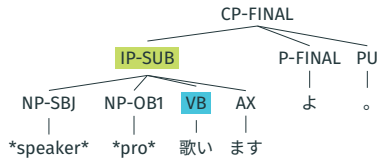
複雑な関係表現：「または」

パイプ（|）を２つの関係表現の間に置くと、「または（そのうちの一方が満たされる）」の意味になる。例えば：

```
VB > /^IP-MAT\b/ | > (/^IP-SUB\b/ > /^CP-FINAL\b/)
```

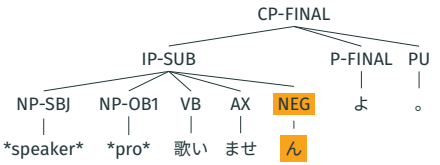
> 上記は、動詞を述語とする文のうち、いわゆる平叙文（終助詞を伴うものを含む）を抽出する。

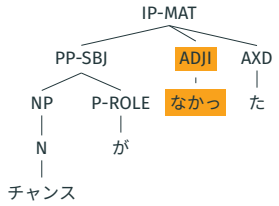




複雑な関係表現：「または」

否定文の抽出は、タグ NEG を使っただけでは不十分。動詞「ある」に対応する「ない」は形容詞（ADJI）扱い。





このようなときも、パイプ（|）を使うと良い。

```
/^IP/ < NEG | < (ADJI < /^(ない|なかっ|なく|なくて|なけれ|なさ|無)/)
```

セッション5の内容

1. いろいろな関係結合子
2. 複雑な関係表現
3. **調べたい表現のアノテーションを知るには**
4. その他
   Tregex と TGrep2 はどう違うの？
   NPCMJ の全データをダウンロードするには

調べたい表現のアノテーションを知るには

調べたい表現の分節のされ方やタグ付けのされ方が分からないと、検索式が書けない！

具体的な表現を文字列検索で検索し、検索結果のツリーを見る

- 文字列検索は、NPCMJ Development Interfaces および NPCMJ Explorer で行うことができる。

新たなウェブサイトに、検索式の例が紹介されている。

- `https://kainoki.github.io/`（検索パターの例）

新たなウェブサイトの Parsing Guide でアノテーション方針の詳細や、具体例を知る。

セッション5の内容

1. いろいろな関係結合子
2. 複雑な関係表現
3. 調べたい表現のアノテーションを知るには
4. **その他**
   Tregex と TGrep2 はどう違うの？
   NPCMJ の全データをダウンロードするには

その他：Tregex と TGrep2 はどう違うの？

- 検索式の書き方は両ツールともほぼ同じ
- TGrep2 (Rohde 2005)
- Douglas Roland 氏が実行可能なものを公開してくれている。
- `http://www.acsu.buffalo.edu/~droland/tgrep2/`
- 括弧付きツリー形式のファイルから検索用のファイルを生成することで、検索が可能になる。

その他：Tregex と TGrep2 はどう違うの？

- Tregex (Levy, and Andrew 2006)
- The Stanford Natural Language Processing Group により公開
- 括弧付きツリー形式のファイルに直接アクセス
- verb+https://nlp.stanford.edu/software/tregex.shtml+
- Java プログラム
- コマンドプロンプトから実行できるほか、GUI のインターフェース（TregexGUI）もある（最新版の zip ファイルをダウンロードするとよい）。

その他：NPCMJ の全データをダウンロードして使いたい

- NPCMJ のウェブサイトから今年度 “正式” 公開データがダウンロードできる（括弧付きツリー形式、90069 ツリー、1304508 語）
- `https://npcmj.ninjal.ac.jp/`
- 国語研でのプロジェクト終了後、現在の NPCMJ は Kainoki コーパスと呼ばれることになる。データー式は GitHub で公開される予定
- `https://kainoki.github.io/`

その他：NPCMJ の全データをダウンロードして使いたい

TregexGUI