
http://npcmj.ninjal.ac.jp/

# 統語・意味解析情報付き日本語コーパスの構築に向けて

# (2) アノテーション方式とコーパスの特色

吉本　　啓
東北大学

# 1　はじめに

- アノテーション方式
  文法研究のための、テクストデータのピンポイントの検索
  意味解析情報の抽出
- アノテーションの原則
  Penn Historical Treebank に従う
  日本語実情に適合させる
- コーパスを XML でエンコーディング
  インターネットを介した高速、効率的利用
- アノテーション支援ツール
  Emacs 利用
- 特色と意義

# 2 アノテーションの方式 (1)

従来の文節コーパス

検索結果からの手作業による選択が必要

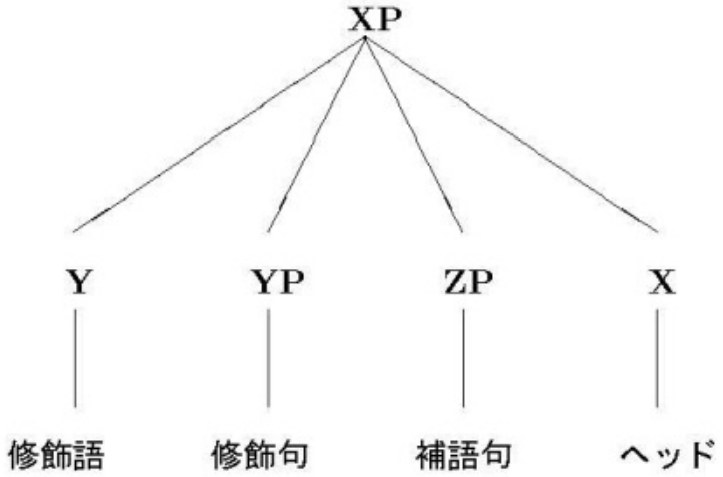
必要十分な検索結果を返してくれるコーパスは作れないか？

- 統語構造 (句構造) のアノテーション
- 意味解析により、複雑な構文における語句間の関係も把握

# 2 アノテーションの方式 (2)

すべての種類の句が同一の比較的フラットな構造*

- 検索や意味解析を目的とする
  木構造の処理が容易
- スコープ (作用域) 包含
  関係の指定にあたって、
  統語的埋め込みによる
  干渉を防ぐ




*2分木と同等、自動変換可

様々な言語理論に対して中立的

# 2 アノテーションの方式 (3)

- ペン通時コーパス (Penn Historical Corpora) の方式を採用

  世界の多様な言語のコーパスに利用

  英語、フランス語、アイスランド語、ポルトガル語、ギリシャ語、イディッシュ語等

- 世界からの利用が容易

- 他の言語のコーパスとの比較、対照が容易

# 2　アノテーションの方式 (4)

句・節の機能のタグ付けにより、より正確な統語情報を提供

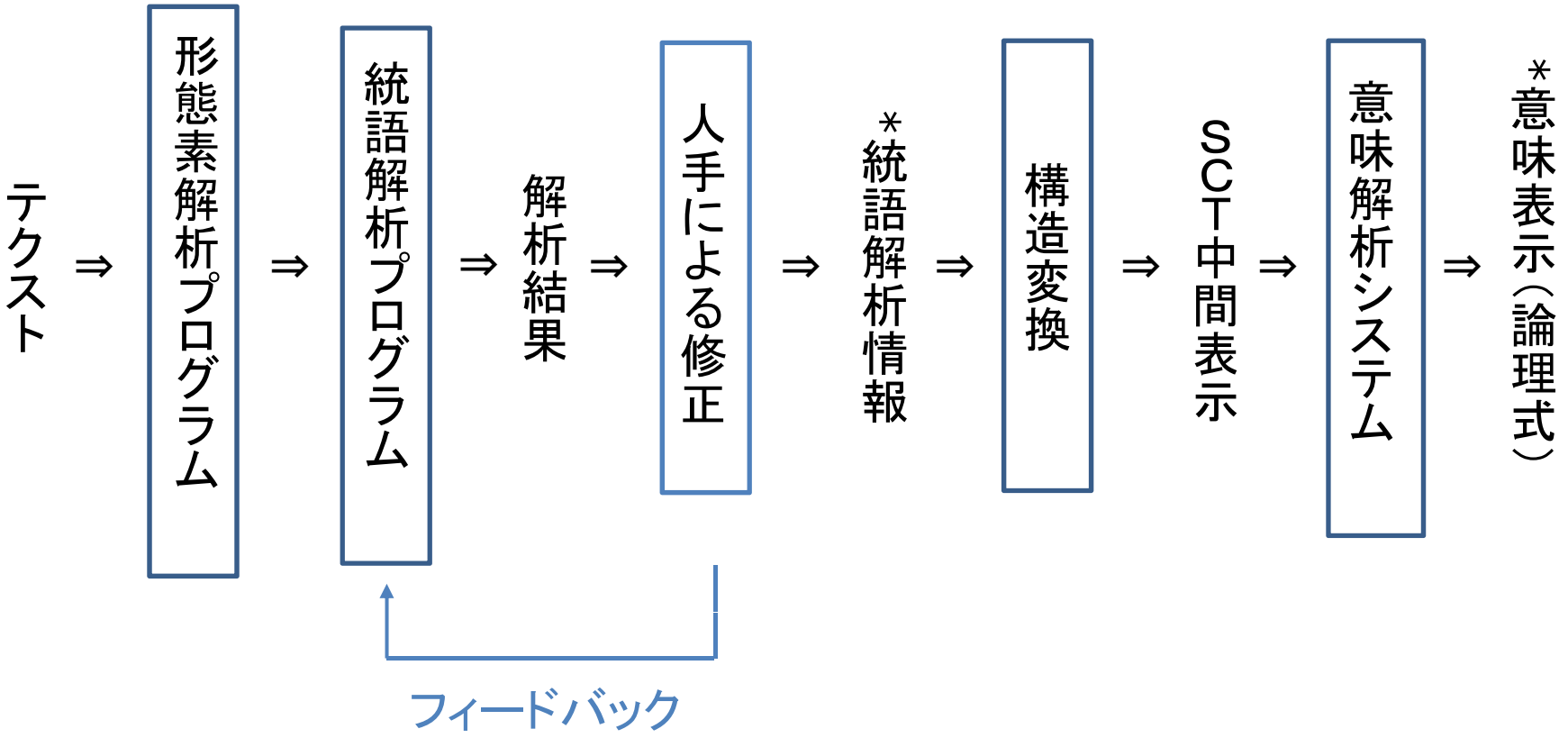
統語構造の曖昧性を克服、意味情報を抽出

統語ラベル-機能ラベル

例　PP-SBJ, PP-OB1, PP-OB2

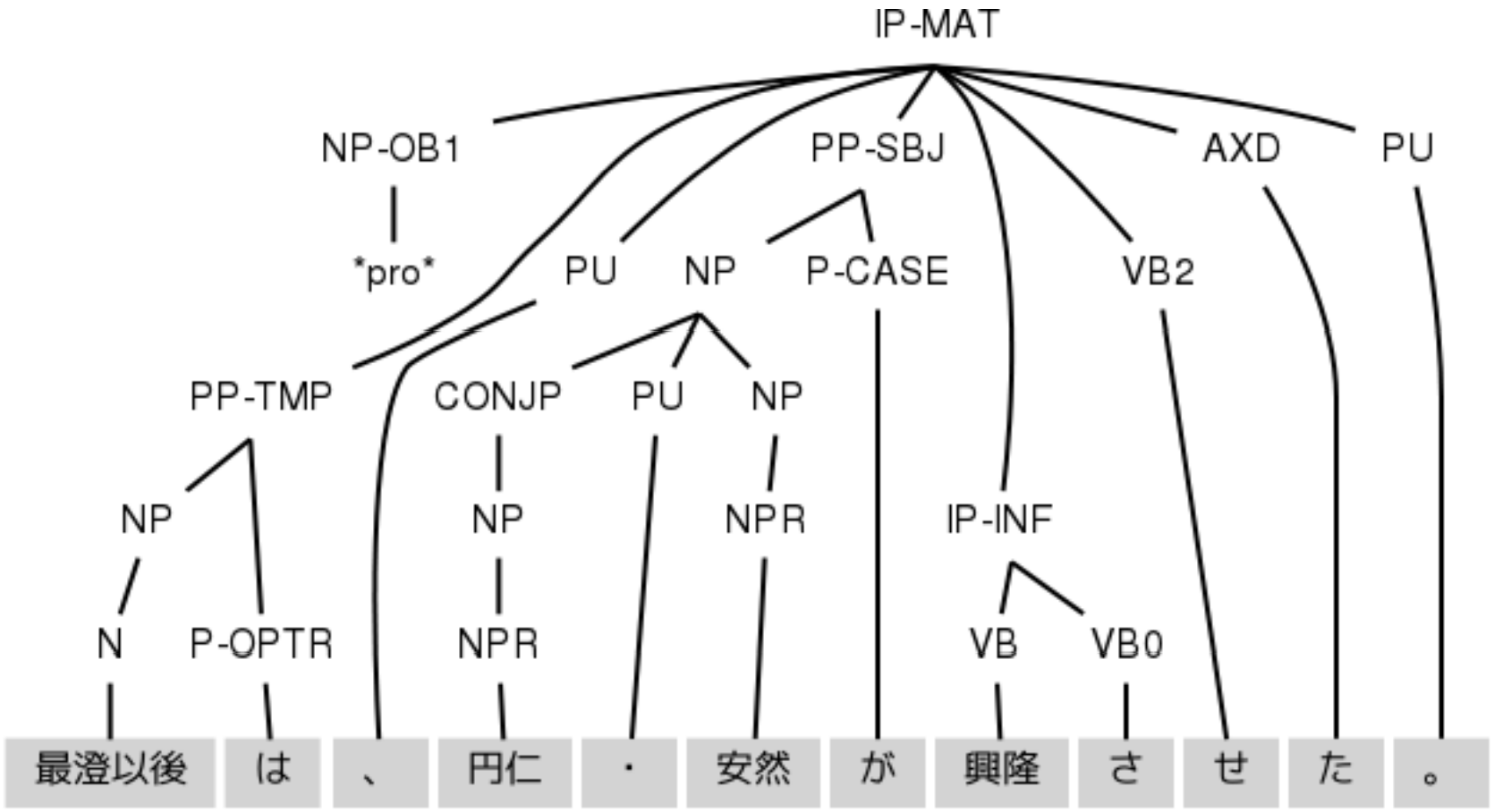
IP-REL, IP-EMB

# 3　構築方法 (1)




*統語・意味解析情報付き現代日本語コーパス

# 3 構築方法 (2)


IP-MAT
NP-OB1
PP-SBJ
AXD
PU
*pro*
PU
NP
P-CASE
VB2
PP-TMP
CONJP
PU
NP
NP
NP
NPR
IP-INF
N
P-OPTR
NPR
VB
VB0
最澄以後
は
、
円仁
・
安然
が
興隆
さ
せ
た
。

# ４　ラベル (1)

語彙ラベル・・・27種類 + 記号

| | |
|---|---|
| ADJI | い-形容詞 |
| ADJN | な-形容詞 |
| ADJT | たる-形容詞 |
| ADV | 副詞 |
| AX | 助動詞 |
| AXD | 助動詞「た」 |
| CARD | 数詞 |
| CONJ | 並列接続詞 |
| D | 限定詞 |
| FW | 外国語 |
| INTJ | 間投詞 |
| MD | モーダル助動詞 |
| N | 普通名詞 |
| NEG | 否定辞 |
| NPR | 固有名詞 |
| NUMCL | 助数詞 |
| P | 助詞 |
| PASS | 受動助動詞 |
| PRO | 代名詞 |
| Q | 量化詞 |
| VB | 動詞 |
| VB0 | 軽動詞 |
| VB2 | 補助動詞 |
| WADV | 疑問副詞 |
| WCARD | 疑問数詞 |
| WD | 疑問限定詞 |
| WPRO | 疑問代名詞 |

# 4　ラベル (2)

句のラベル・・・13種類

| | | | |
|---|---|---|---|
| ADJP | 形容詞句 | NUMCLP | 数助詞句 |
| ADVP | 副詞句 | PP* | 後置詞句 |
| CONJP | 接続詞句 | PRN | カッコ挿入句 |
| CP* | 節 | QP | 量化詞句 |
| FRAG | 断片 | | |
| INTJP | 間投詞句 | | |
| IP* | 節 | | |
| NML | 中間名詞節 | | |
| NP* | 名詞句 | | |

＊・・・　機能タグを付加できる

# 5　アノテーションの原則

- 一貫性
- 検索利用の容易性
- 意味解析のための適合性
- 客観性
- 言語使用の実態の反映

# 6　アノテーションの特徴 (1)

1つの機能語として働く連語は、1つの助詞 (P) として扱う

として
について
に対して/対する
に関して/関する

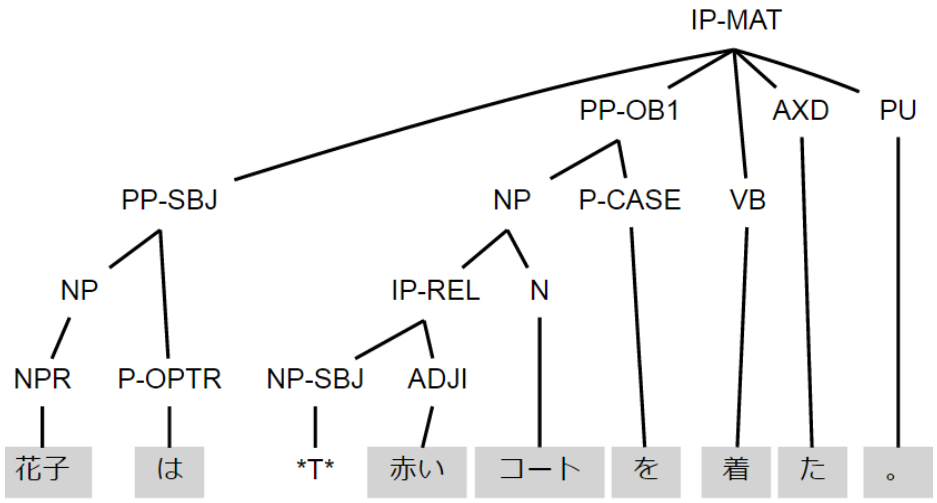
# 6　アノテーションの特徴 (2)

1つのモーダル的機能を果たす連語は、1つの助動詞 (MD) として扱う

かもしれない
ざるをえない
てはならない
てもよい
なくてはならない
なければならない
わけにはいかない

# 6 アノテーションの特徴 (3)

主語や目的語として機能する後置詞句 (PP) には文法機能 (SBJ, OB1, OB2) を追加



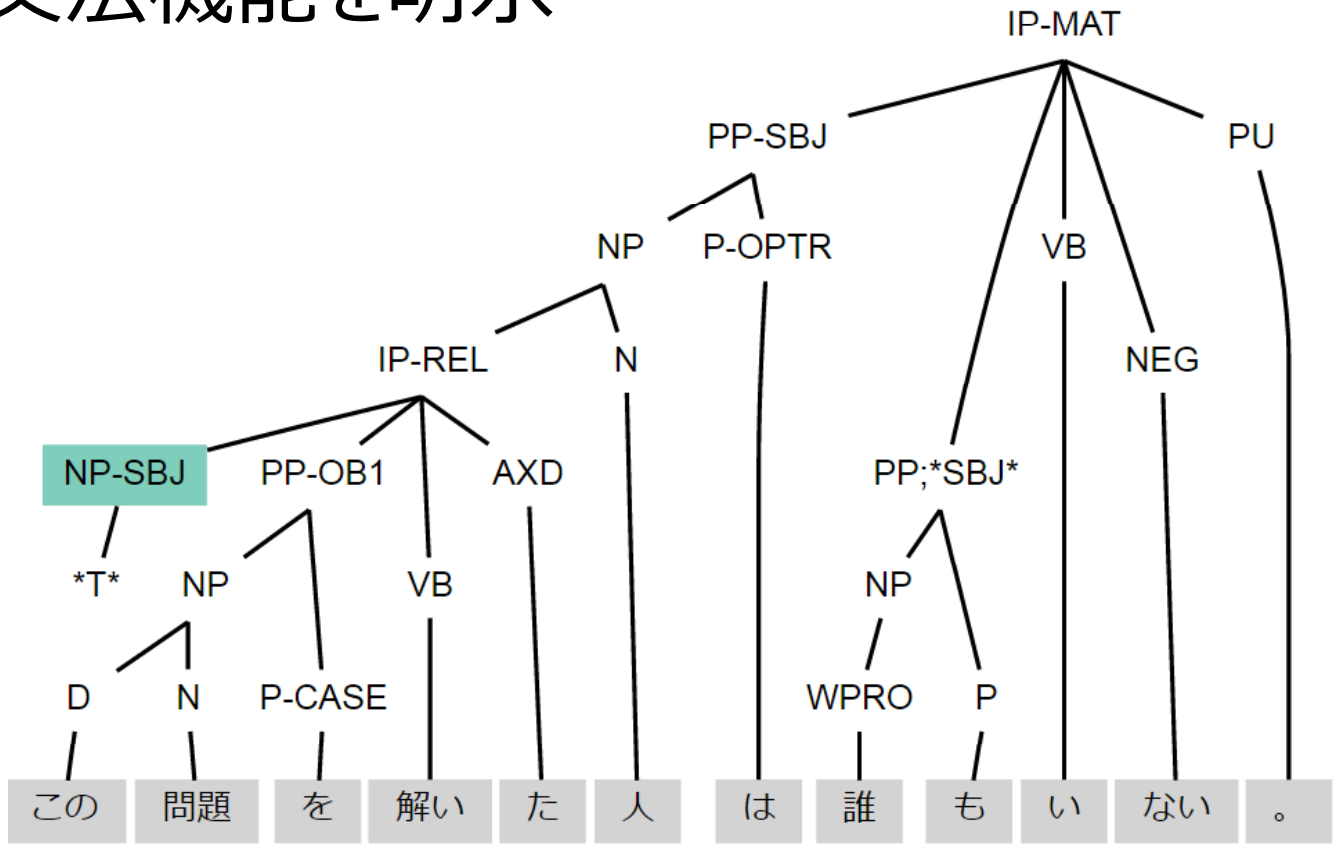

フロントエンド

```
( (IP-MAT (PP (NP (NPR 花子))
          (P は))
        (NP-SBJ *)
        (PP (NP (IP-REL (NP-SBJ *T*)
                        (ADJI 赤い))
                (N コート))
            (P を))
        (NP-OB1 *を*)
        (VB 着)
        (AXD た)
        (PU 。))
(ID 10_textbook_kisonihongo;page_13;AT1-10;MJ))
```

バックエンド

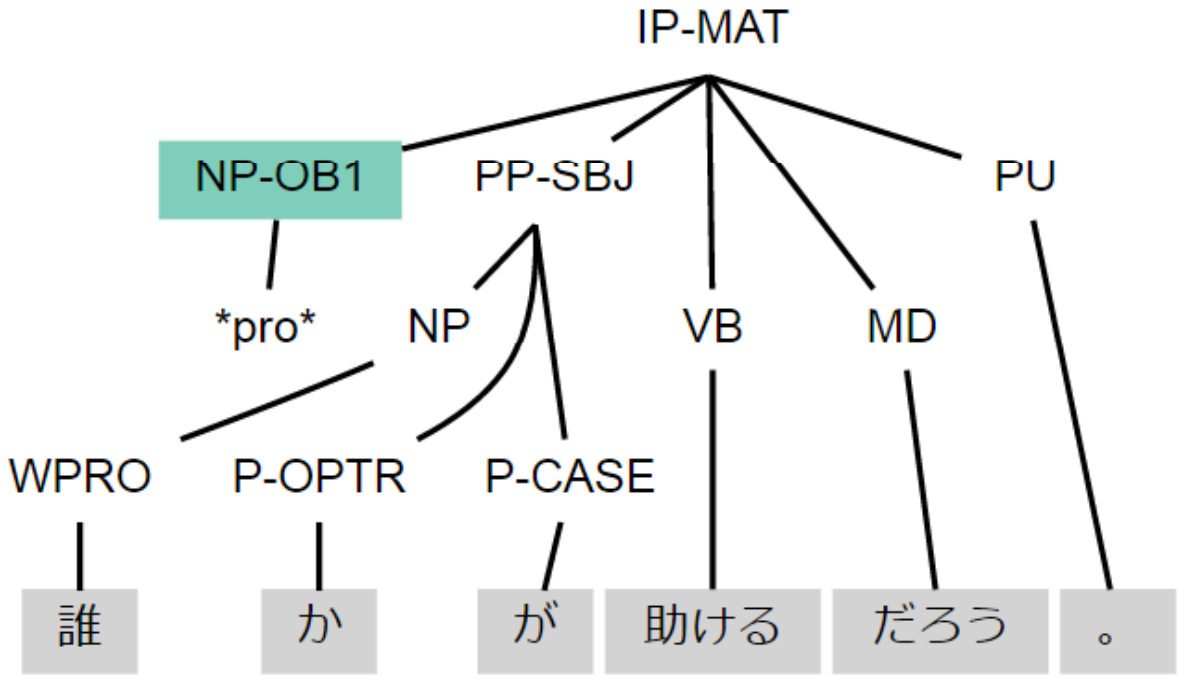
# 6 アノテーションの特徴 (4)

関係節内に空所 (トレース) に相当するノードを与えて文法機能を明示



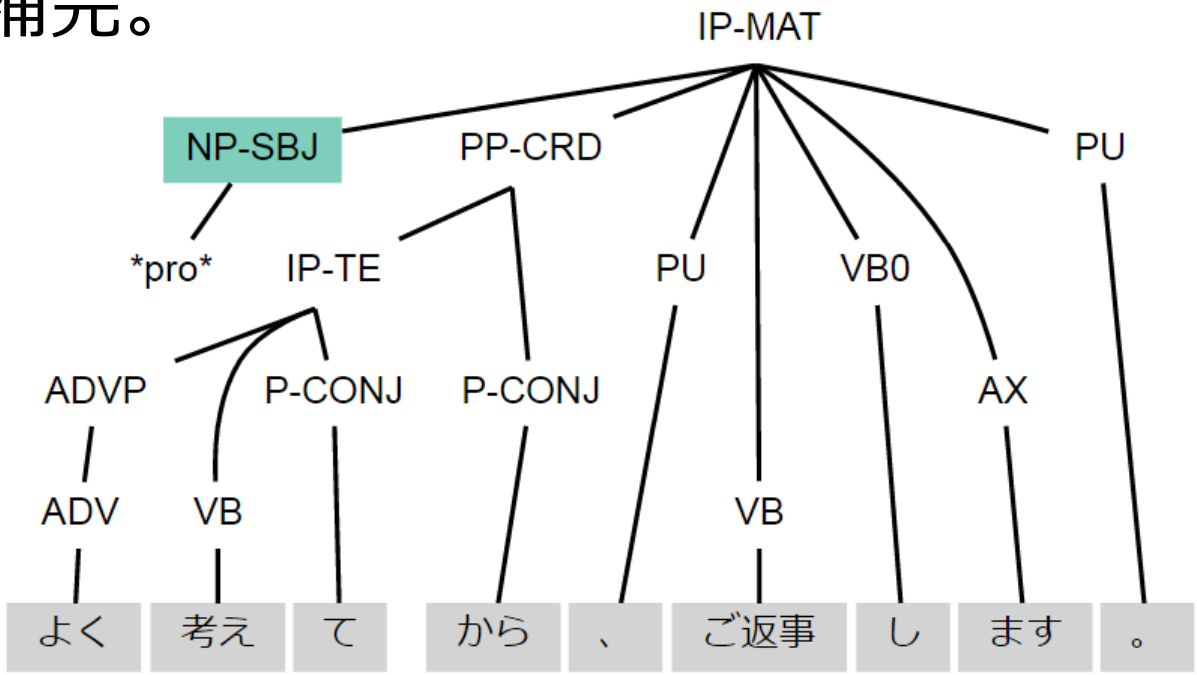
# 6　アノテーションの特徴 (5)

ゼロ代名詞を明示



# 6 アノテーションの特徴 (6)

埋め込まれた用言の主語/目的語が主節の主語/目的語によりコントロールされている場合は、ゼロ代名詞としてタグ付けしない (cf. 南 1974 の文階層構造)。意味解析により情報を補完。



# 6 アノテーションの特徴 (7)

インデクスは使用しない
　意味解析により情報を補完
　アノテーションの作業量を軽減

例外:
　　外置
　　数量詞遊離
　　主要部内在型関係節

# 7　XML によるエンコーディング (1)

XML ・・・ 文章やデータの意味や構造を記述するための汎用的マークアップ言語

- 効率的な検索
- ノード情報の拡張 (レンマなど) が可能
- ウェブインタフェース技術や XML によるアノテーションの操作・検索用ツールを利用できる
- これまでのカッコ表示を XML マークアップ方式に自動変換

# 7　XML によるエンコーディング (2)

カッコを用いたアノテーション

　輪を回す

```
( (IP-MAT (NP-SBJ *pro*)
        (PP (NP (N 輪))
            (P を))
        (NP-OB1 *を*)
        (VB 回す)
        (PU 。))
  (ID 2_textbook_kisonihongo;page_13;AT1-2;MJ))
```

# 7 XML によるエンコーディング (3)

## XML マークアップ方式

```
<alpino ds id="2 textbook kisonihongo;page 13" version="1.3">
  <node cat="ip-mat" id="1" begin="0" end="5">
    <node cat="np-sbj" id="2" begin="0" end="1">
      <node pt="zero" word="*pro*" id="3" begin="0" end="1"/>
    </node>
    <node cat="pp-ob1" id="4" begin="1" end="3">
      <node cat="np" id="5" begin="1" end="2">
        <node pt="n" word="輪" id="6" begin="1" end="2"/>
      </node>
      <node pt="p-case" word="を" id="7" begin="2" end="3"/>
    </node>
    <node pt="vb" word="回す" id="8" begin="3" end="4"/>
    <node pt="pu" word="。" id="9" begin="4" end="5"/>
  </node>
  <sentence>*pro* 輪を回す。</sentence>
</alpino ds>
```

# 8　Emacs を利用したアノテーション (1)

Emacs けやきモード

アノテーターによる文解析結果修正作業支援ツール

Emacs エディターのマクロ機能を利用

マウスによるメニュー操作での修正

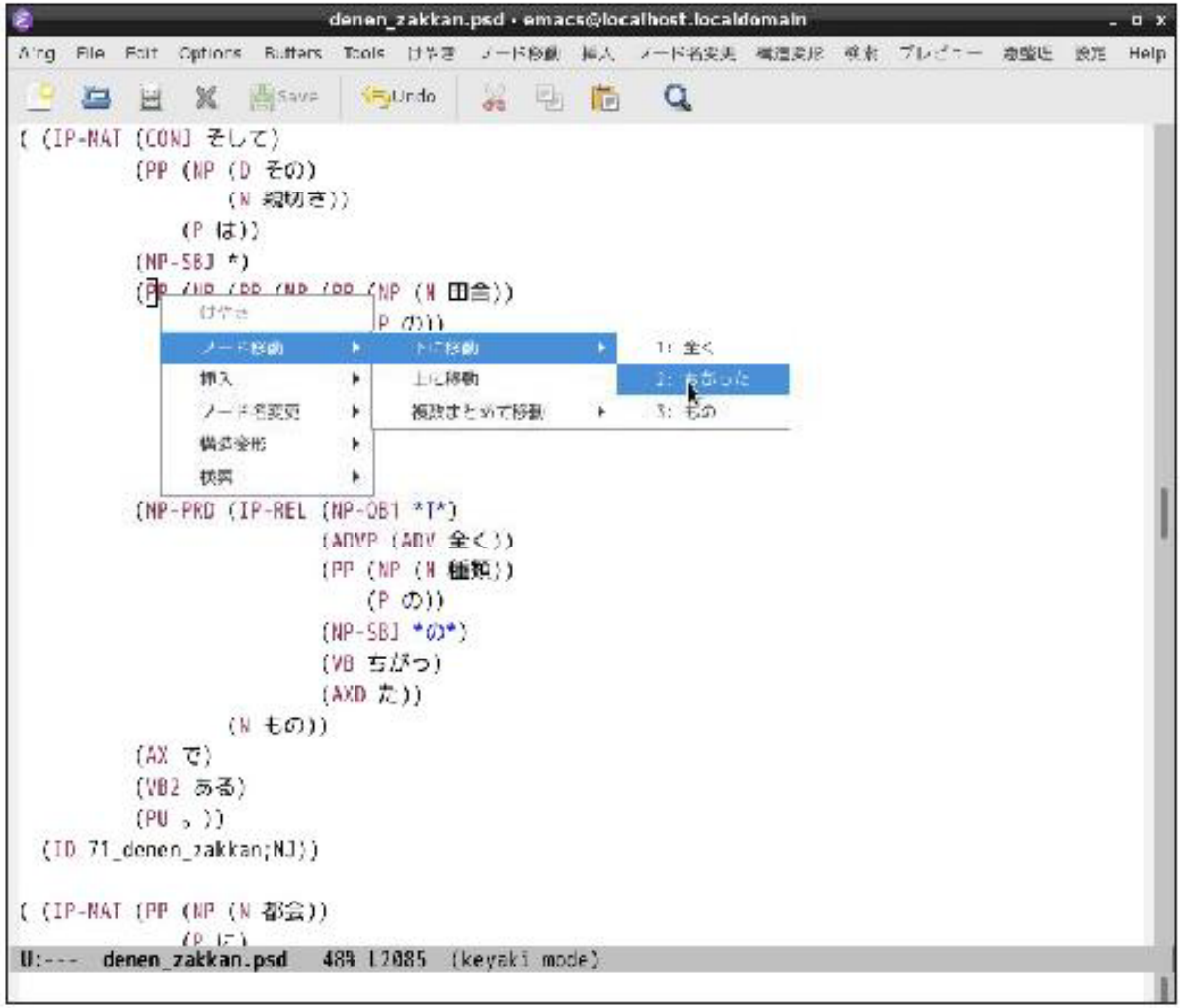
- ノードの係り先

  「係り先の述語は何か」という言語学的問題に集中できる

- ノードラベル

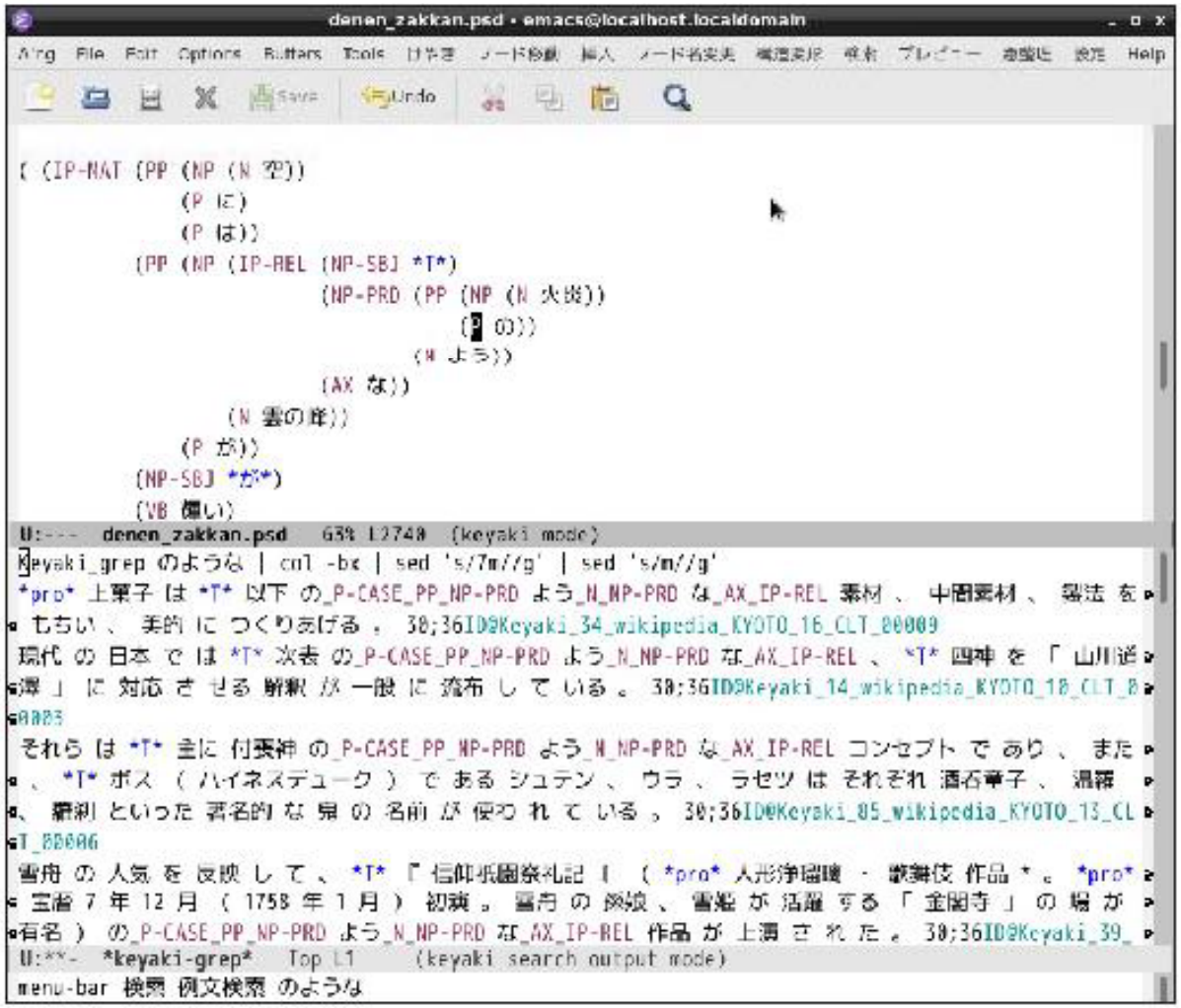
# 8　Emacs を利用したアノテーション (2)

修正作業
の画面



# 8 Emacs を利用したアノテーション (3)

修正済み
コーパス
からの検索



# ９　まとめー特色と意義

従来のコーパス

　語句間の共起 (co-occurrence) に関する手掛かりを与えるだけ　⇒　手作業が必要

本ツリーバンク

　Unbounded dependency など複雑な構文も含め、語句間の依存関係 (dependency) を把握

研究者が必要とする文法情報をピンポイントで得られる