現代日本語統語・意味解析コーパス（NPCMJ）の概要と検索 †

プラシャント・パルデシ（国立国語研究所） 長崎郁（名古屋大学） 鈴木彩香（国立国語研究所）

NPCMJとは

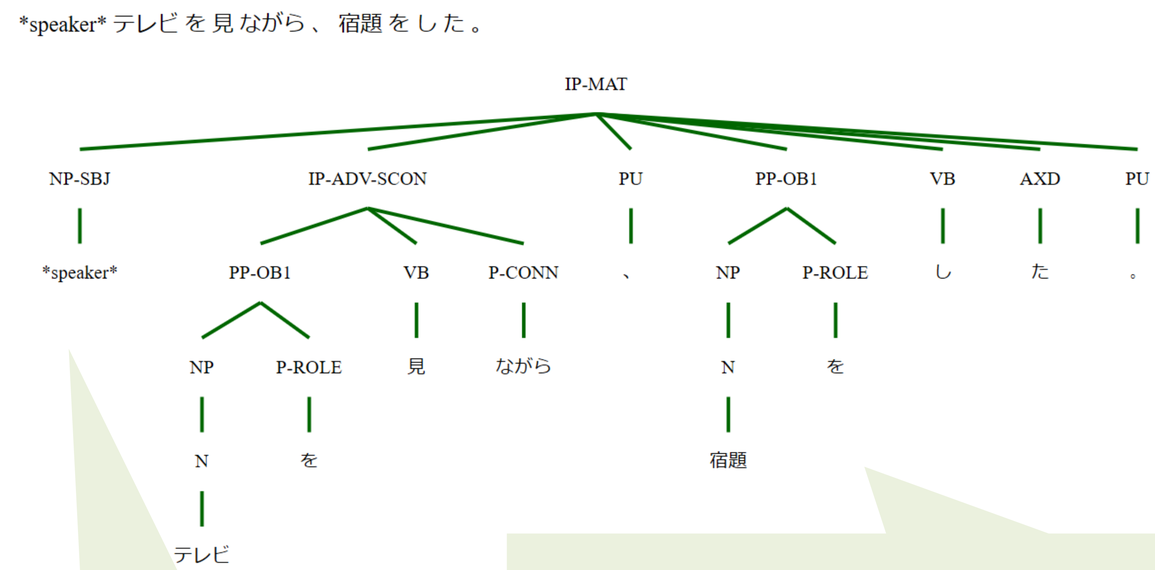
►NINJAL Parsed Corpus of Modern Japanese (NPCMJ)

- ► **文の構造や意味に関する詳細な情報**が付与された言語資料
  - ► 日本語の**様々なタイプの文**を検索し、言語研究に役立てることができる
- ► 国立国語研究所で**2016年4月**から開発開始
  - ► http://npcmj.ninjal.ac.jp/
- ► 上記ウェブサイトから**全文データのダウンロード・各検索インターフェース**が利用できる
  - ► 2020年3月現在、**4万文**を無償公開中
  - ► 毎年1万文追加、プロジェクト終了（2022年3月）までに**6万文**公開予定

データの出典（2020年3月時点）

出典	文数
青空文庫	9,561
聖書	1,664
その他の小説の抜粋	923
ノンフィクション	223
書籍の一部	553
ウィキペディア	2,556
新聞記事	4,777
法律文	337
国会議事録	1,698
テッドトーク	1,453
教科書	6,953
辞書	5,362
その他	2,389
合計	**40,831**

統語分析

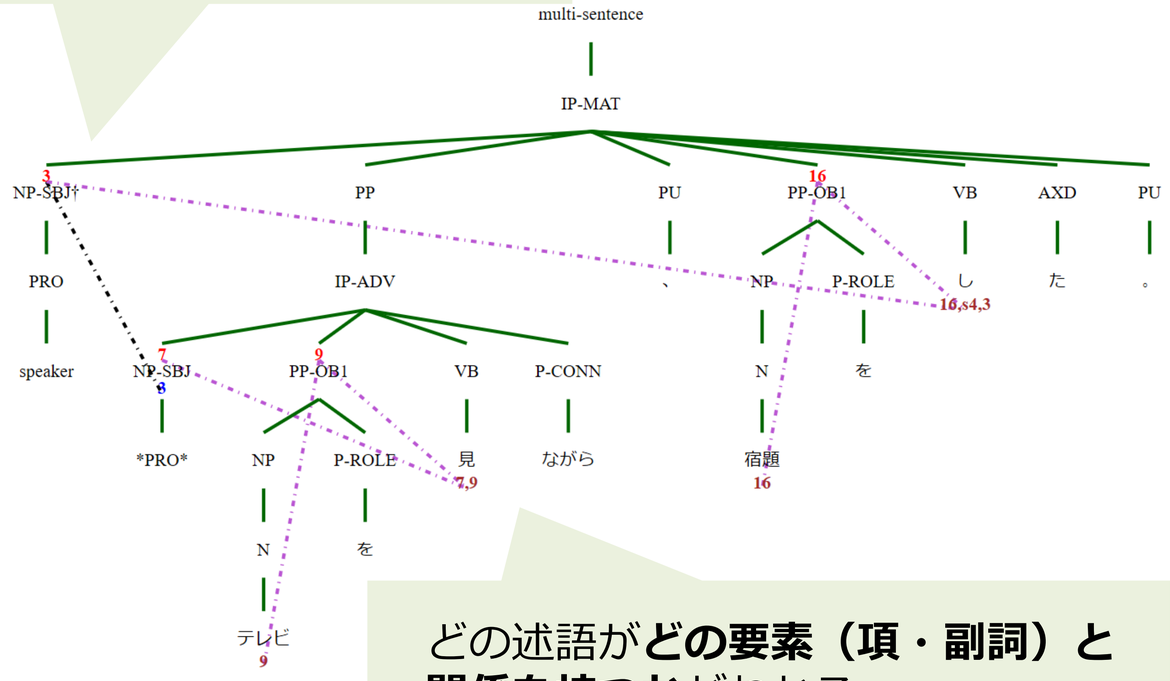


省略された主語や目的語もわかるようになっている

品詞、句、それらの修飾関係や、**主語・目的語といった句のはたらき**がわかる

意味解析

名詞が持つ指示がどのように受けつがれるかがわかる



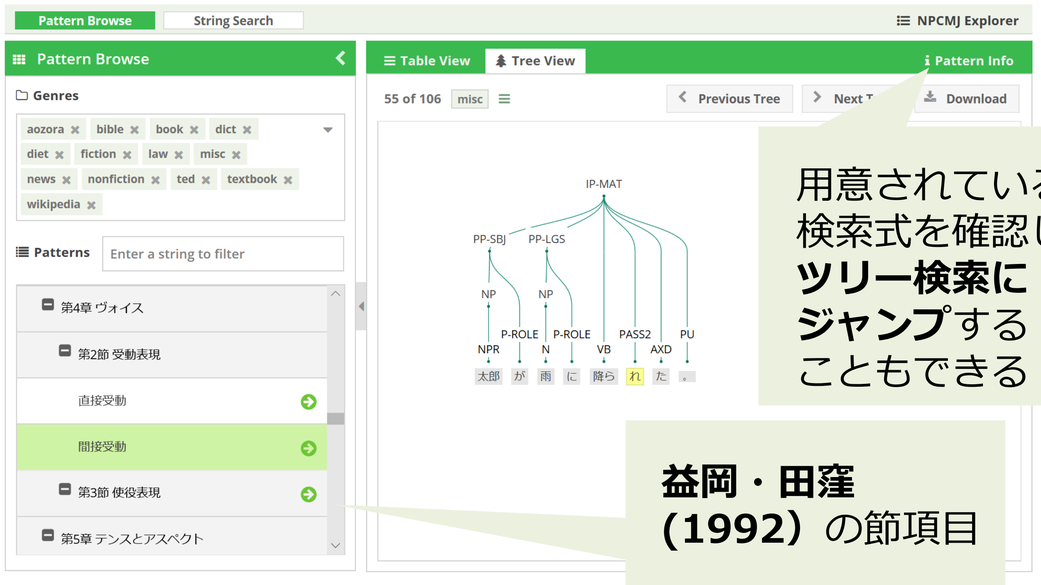
どの述語が**どの要素（項・副詞）と関係を持つか**がわかる

検索インターフェース

初中級者向け

►NPCMJ Explorer　Pattern Browse

- ► **日本語のいくつかの主要な文法項目**に対して、それを調べるための検索式が用意されている
- ► 項目名をクリックするだけでコーパス中の用例を見ることができる（http://npcmj.ninjal.ac.jp/explorer/）



用意されている検索式を確認し、**ツリー検索にジャンプ**することもできる

益岡・田窪（1992）の節項目

中上級者向け

►NPCMJ Search　ツリー検索

- ► 自分で検索式を書いて、コーパス中の用例を調べることができる
  - ► たとえば…
    「の」でマークされた主語

TGrep-lite 検索結果

```
[PP-SBJ] < ([P-ROLE] < の)
```

Submit　all data　□ reveal

1. それ は 全く 前例 の ない こと で あっ た が 、 *T* 医学 に *T* 深い 理解 を いた は 、 これ を 許し た 。 book_excerpt-44_7
2. *pro* あなた も よく ご存じ の よう に 、 *T* ほとんど 一 年じゅう 店 の 外 に い の セールスマン は 、 かげ口 や 偶然 や *T* いわれ の ない 苦情 の 犠牲 に なり そうした もの を 防ぐ こと は まったく でき ない ん です 。 aozora_Harada-1960_374
3. 地 の ある 限り 、 *speaker* 種まき の 時 も 、 刈入れ の 時 も 、 暑さ 寒さ も 、 、 昼 も 夜 も やむ こと は ない であろう 」 。 bible_old_242

『主語となる助詞句（PP-SBJ）が、「の」を支配する格助詞（P-ROLE）を支配している』

►**文字列検索**や、**検索式の作成を補助**するインターフェースなどもある（http://npcmj.ninjal.ac.jp/interfaces/）

参考文献：益岡隆志・田窪行則(1992)『基礎日本語文法』（くろしお出版）
† 本研究は国立国語研究所「統語・意味解析コーパスと言語研究」プロジェクトの支援を受けています