関西言語学会第４４回大会
シンポジウム

高度文法情報付きコーパスと
その日本語研究への応用

2019年7月14日
関西大学　千里山キャンパス

シンポジウムの構成

- 第I部: 概要とウェブ・インターフェースを用いた検索の例
  1. プラシャント-パルデシ (国立国語研究所)・吉本啓 (東北大学)
  「イントロダクション」
  2. 窪田悠介 (国立国語研究所)・峯島宏次(お茶の水女子大学)
  「前提投射の統語コーパスでの検索」
- 第II部: 統語コーパスを用いた文法研究
  3. 三好伸芳 (実践女子大学)
  「名詞句と述語の共起関係から見たコーパス研究」
  4. 井戸美里 (国立国語研究所)
  「述語名詞句の中の連体修飾節」
- 第III部: プログラミングを伴う言語研究
  5. 大久保弥 (東京外国語大学・国立国語研究所)
  「並列構文における主語句標識と文の意味解釈」

「イントロダクション」

プラシャント-パルデシ (国立国語研究所)・
吉本啓 (東北大学)

日本語の特徴

- 項の省略
- あいつはそこにいた。聞かれていたかもしれない。

- 名詞修飾表現
- [[子どもが見ている]写真]（「内」の関係）
- [[子どもが泳いでいる]写真]（「外」の関係）

大量のデータにこのような情報を付加したコーパスがあればいいな。

X

- あいつはそこにいた。
- （Yが） （Xに） 聞かれていたかもしれない。

文法的な主語　論理的な主語　受け身　モダリティー

5

しかし。。。。

- 3年半前にそのようなコーパスは存在しなかった。
- 現在は、公開されている。
- それはNPCMJである。
- 本日のセッションの主役はNPCMJである。

6

NPCMJとは

- NINJAL Parsed Corpus of Modern Japanese
- 統語・意味解析情報付き現代日本語コーパス
- http://npcmj.ninjal.ac.jp
- 2019年3月現在：30460文・ツリーを公開
- 2022年3月までに：6万文・ツリーを公開する予定

7

公開可能な、様々なジャンルのデータを選定し、統語・意味情報を付加した上で無償公開

8

NPCMJ検索用のインターフェース（検索ツール）

9

初中級者向け検索ツール
NPCMJ Explorer

10

初中級者向け検索ツール
NPCMJ Explorer

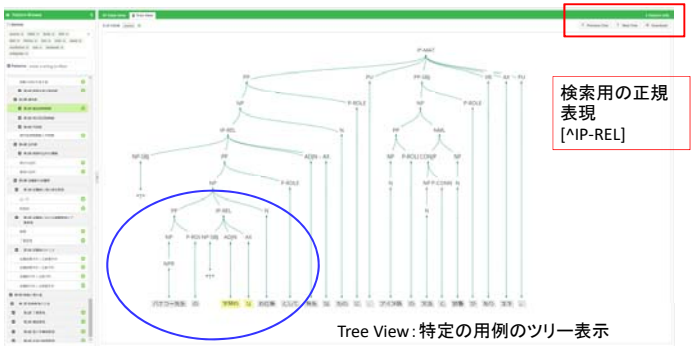
益岡隆志・田窪行則著『基礎日本語文法―改訂版―』（くろしお出版）第Ⅳ部　複文
第3章連体節第2節補足語修飾節（公開中の約3万文中14548文、いわゆる「内の関係」）

Table View：検索用例の一覧

11

初中級者向け検索ツール
NPCMJ Explorer

益岡隆志・田窪行則著『基礎日本語文法―改訂版―』（くろしお出版）第Ⅳ部　複文
第3章連体節第2節補足語修飾節（公開中の約3万文中14548文、いわゆる「内の関係」）



Tree View：特定の用例のツリー表示

初中級者向け検索ツール
NPCMJ Explorer

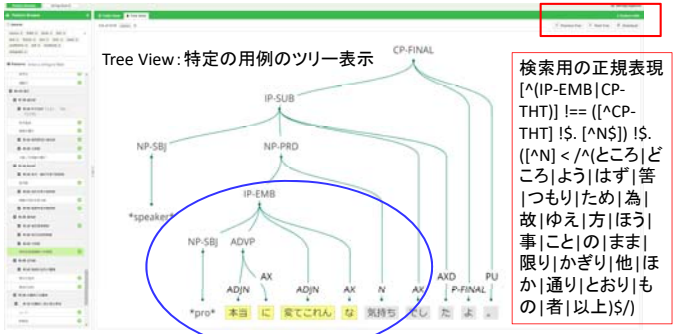
益岡隆志・田窪行則著『基礎日本語文法―改訂版―』(くろしお出版)第Ⅳ部　複文
第3章連体節第4節内容節(公開中の約3万文中3179文、いわゆる「外の関係」)

Table View：検索用例の一覧

13

初中級者向け検索ツール
NPCMJ Explorer

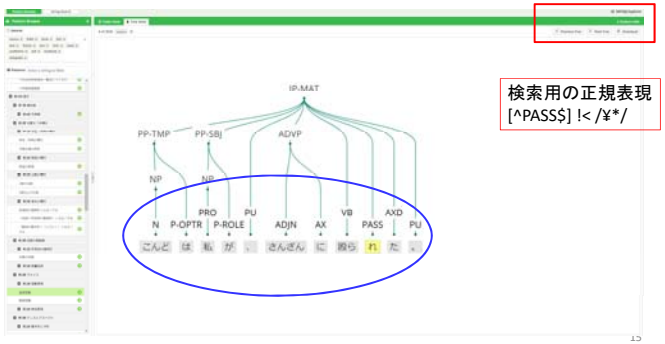
益岡隆志・田窪行則著『基礎日本語文法―改訂版―』(くろしお出版)第Ⅳ部　複文
第3章連体節第4節内容節(公開中の約3万文中3179文、いわゆる「外の関係」)



14

初中級者向け検索ツール
NPCMJ Explorer

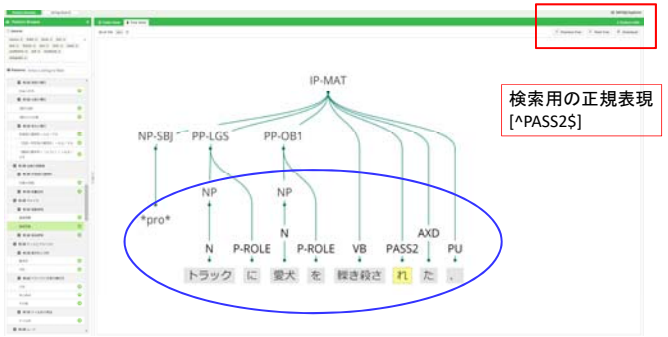
益岡隆志・田窪行則著『基礎日本語文法―改訂版―』(くろしお出版)第Ⅲ部　単文
第4章ヴォイス第2節直接受動(公開中の約3万文中2926文)



15

初中級者向け検索ツール
NPCMJ Explorer

益岡隆志・田窪行則著『基礎日本語文法―改訂版―』(くろしお出版)第Ⅳ部　複文
第3章連体節第4節内容節(公開中の約3万文中3179文、いわゆる「外の関係」)



16

NPCMJ Explorerを本格的に利用するために

- Tgrep-liteという検索言語を学ぶ必要がある
- それほど難しいものではないが、チュートリアルを受ける必要がある
- プロジェクトでは年に2回チュートリアルを開催しているのでぜひご参加ください
- こんなことができるかのような問い合わせはNPCMJのHPの「お問い合わせ」でどうぞ。
  http://npcmj.ninjal.ac.jp/
- よろしくお願いします。

17

NPCMJ開発の動機と特徴

吉本　啓

18

1　なぜツリーバンクか

日本語コーパスの現状

- 単語への分割
- 品詞情報 (名詞、動詞、助詞、助動詞、…)
- 文節にもとづく解析と文節間の係り受け

19

2.1 なぜ句構造か

○付属語が独立した文法機能を果たすことがある
・[冷蔵庫 に] [牛乳 が] [入っ て い ない]
$\exists xy$(冷蔵庫(x) ∧ 牛乳(y) ∧ ¬入っている(y, x))
・[[冷蔵庫に牛乳が入ってい] ない]
$\exists x$ (冷蔵庫(x) ∧ ¬ $\exists y$(牛乳(y) ∧入っている(y, x)))
○ 1つの形態素が多数の文法機能を表す
が ⇔ 主語、目的語
1つの文法機能が多数の形態素により表される
主語 ⇔ が、は、に、無表示 …
○係り受けだけでは統語情報として不十分
- [昨日] [とっ た] [写真] (内の関係)
- [子供 が] [泳い で いる] [写真] (外の関係)

20

2.2 なぜ句構造+論理意味表示か

課題　格フレームの抽出（例: 動詞と目的語名詞）
同一の節の中にあらわれるとは限らない
長距離依存（関係節、主題等）により隔てられる可能性
e.g.　私が買ったメロン
私が買って、冷蔵庫に入れてあるメロン
私が買って、冷蔵庫に入れてあると思ったメロン

- 文節中心のコーパス/従来の統語解析情報付きコーパスでは、共起関係しか分からない
- 論理意味表示　$\exists xy(\ldots \text{メロン}(y) \wedge \text{買う}(x, y))$
→ 依存関係 (dependency) の抽出
→ ピンポイントの検索

21

3 アノテーションの方式 (1)

- ペン通時コーパス (Penn Historical Corpora; Santorini 2010) の方式を採用
世界の多様な言語のコーパスに利用

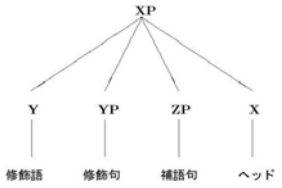
- 統語情報をラベル付きのカッコにより表示
品詞タグ ⇒ 単語
統語タグ ⇒ 句

22

3 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



様々な言語理論に対して中立的

23

3 アノテーションの方式 (3)

句・節の機能のタグ付け (拡張タグ) 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

24

4 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

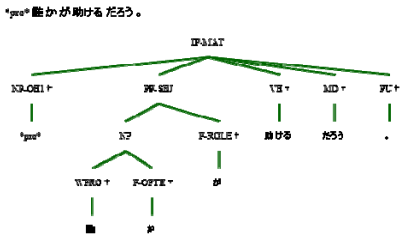
1つのモーダル的機能を果たす連語は、1つの助動詞 (MD) として扱う

> かもしれない、相違ない、ちがいありません、違いない

25

4 アノテーションの特徴 (2)

○ゼロ代名詞を明示



○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

○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。意味解析により情報を補完

26

5 NPCMJ の意義

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
- 長距離依存関係など複雑な構文も含め、語句間の依存関係 (dependency) を把握
  - 関連を持つ語句をすべて相互にリンク付け
- 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

27

ありがとうございました。

謝辞:

本研究はJSPS科研費 JP15H03210および国立国語研究所共同研究プロジェクト「統語・意味解析コーパスと言語研究」の研究成果の一部である。

28