統語・意味解析コーパスの開発と言語研究

Development of and Linguistic Research with a Parsed Corpus of Japanese

NPCMJ コーパスの言語研究への活用: 展望と課題

井戸 美里　(国語研)

鈴木 彩香　(国語研)

窪田 悠介　(筑波大学)

アウトライン

目的:
国立国語研究所で開発中のツリーバンク、NPCMJ **コーパス**
(NINJAL Parsed Corpus of Modern Japanese) **の言語研究への活用を考える**

進め方:

- NPCMJ **コーパスの概要**
- **事例研究**
- **まとめと展望**

はじめに

生成文法研究者の言明 (あるいは弁明?) (2016 **年現在**):

> **また、生成文法研究者は、コーパスからは一例も見つからないような事例を、自身が提案する仮説の証拠として使うことも少なくなく、これについての批判を聞くこともある。しかし、これは、生成文法が統語構造を研究していることと、現在のところ統語構造についてのタグを付与されたコーパスが存在しないことから来る必然的な帰結であって、生成文法の問題というより、コーパスの問題といえるであろう。**
>
> (**小川・長野・菊地** 2016, 13)

統語・意味解析情報付き現代日本語コーパス (NPCMJ)

NINJAL Parsed Corpus of Modern Japanese (NPCMJ)

- **国立国語研究所で開発中のツリーバンク**
  (= 統語構造についてのタグを付与されたコーパス)
- **年間 1 万文を web 上で公開** (http://npcmj.ninjal.ac.jp/)
  - **検索インターフェイスも提供**
  - **「`npcmj ninjal`」で Google 検索するとたどり着ける**
- **昨年度末、最初の 1 万文をインターフェイス** (Ver 1) **とともに公開した。**

ツリーバンクと言語研究

ツリーバンクとは

- 統語構造情報を付与したコーパス
- 文の階層的な構造に基づく検索ができるため、統語論や意味論の研究への活用が期待できる

とはいえ

- (歴史言語学を除いては、) 今のところどちらかというと自然言語処理のリソースと考えられがち (cf. Penn Treebank)

NPCMJ の特徴

- 言語学研究を念頭に開発
- ペン通時コーパスにならい、文法関係やゼロ要素など、言語学的に重要な情報を細かく付与している

ツリーバンクと言語研究

ツリーバンクとは

- 統語構造情報を付与したコーパス
- 文の階層的な構造に基づく検索ができるため、統語論や意味論の研究への活用が期待できる

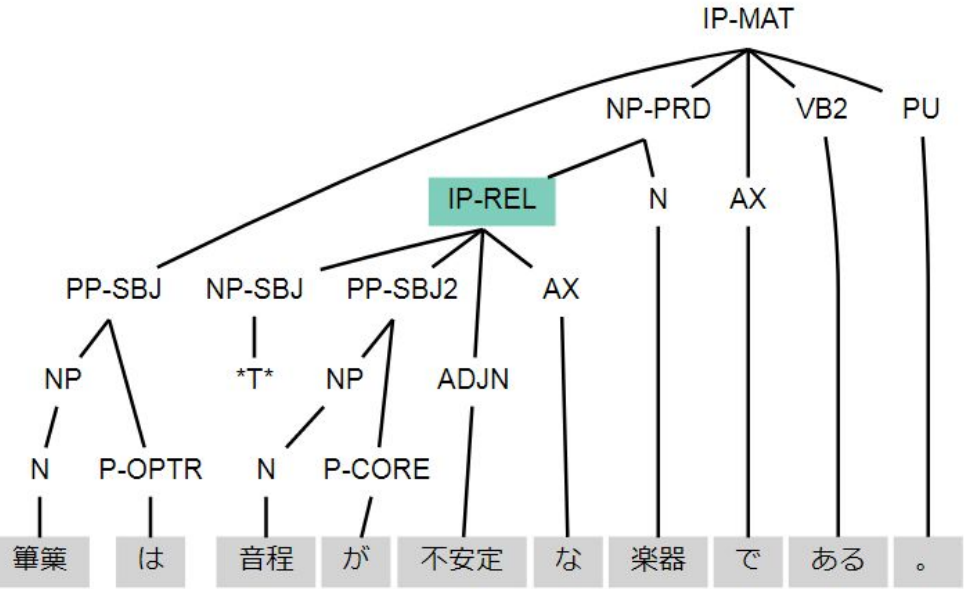
とはいえ

- (歴史言語学を除いては、) 今のところどちらかというと自然言語処理のリソースと考えられがち (cf. Penn Treebank)

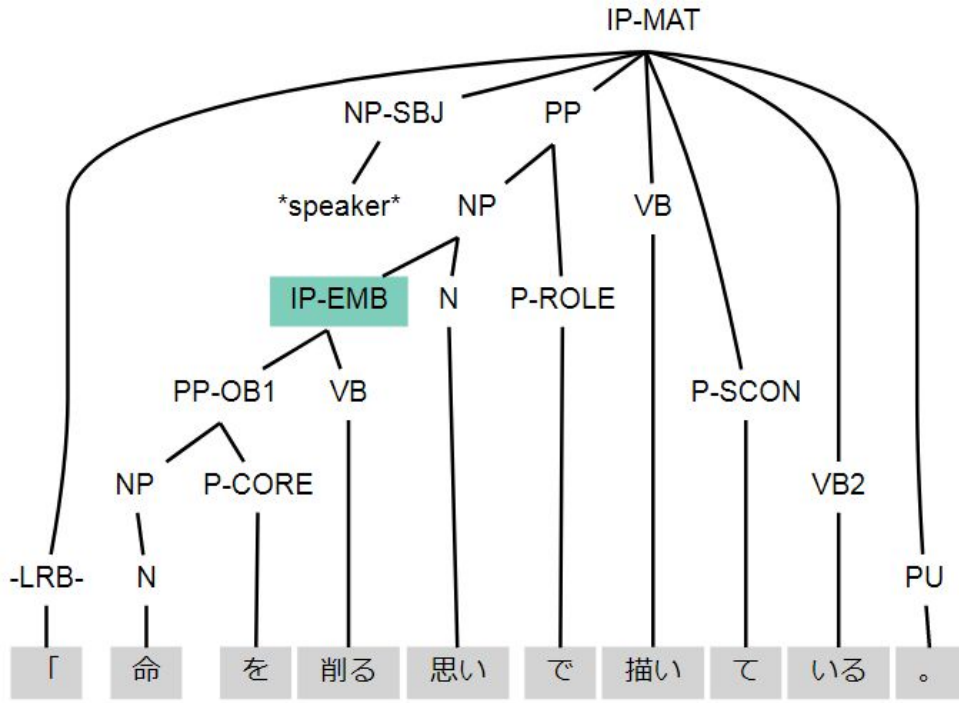
NPCMJ の特徴

- 言語学研究を念頭に開発
- ペン通時コーパスにならい、文法関係やゼロ要素など、言語学的に重要な情報を細かく付与している

ツリーの例



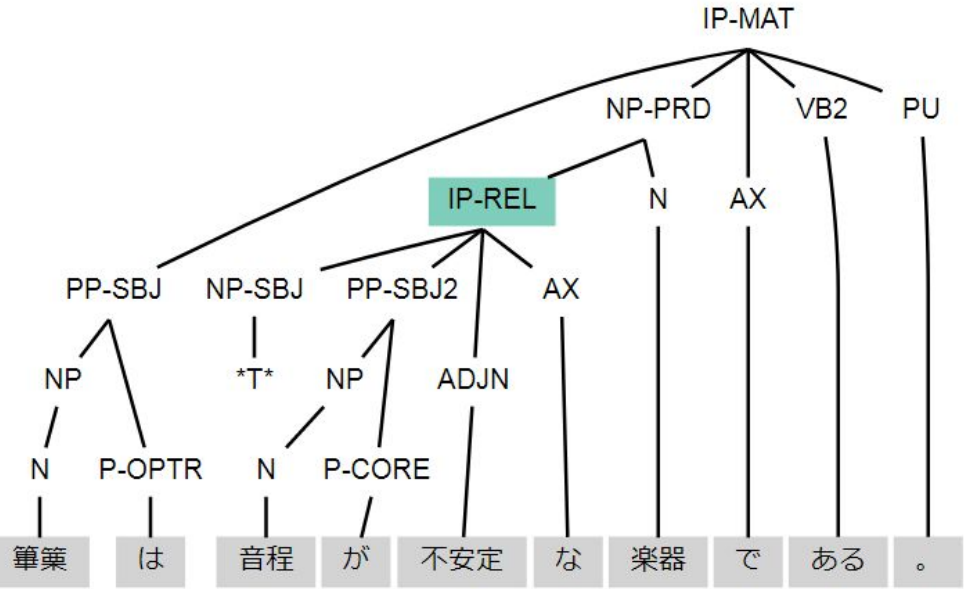
ID: wikipedia_KYOTO_12_CLT_00005_0100



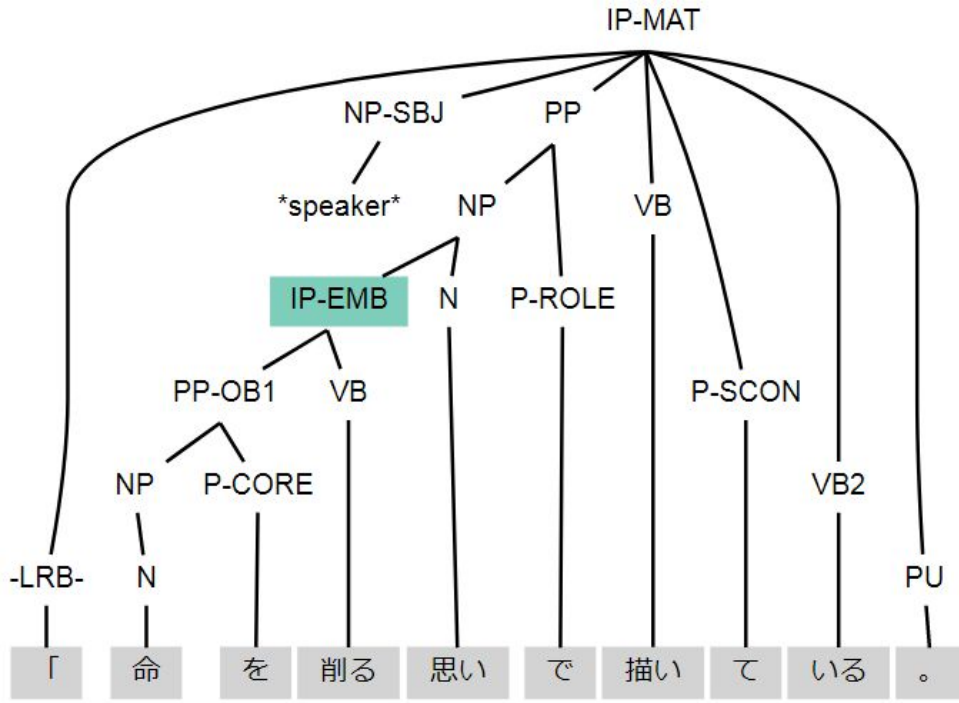
ID: newswire_KAHOKU_00073_K20140101OA0F50XX00001_0082

- 関係節の内の関係と外の関係を区別
- トレースやゼロ主語などの情報も入っている

ツリーの例



ID: wikipedia_KYOTO_12_CLT_00005_0100



ID: newswire_KAHOKU_00073_K20140101OA0F50XX00001_0082

- 関係節の内の関係と外の関係を区別
- トレースやゼロ主語などの情報も入っている

NPCMJ コーパスの内訳

NPCMJ 公式版 (昨年度公開計 1 万文)

テキストジャンル	ツリー数
新聞記事 (河北新報)	4323
Wikipedia 記事	2745
新・旧約聖書	1652
教科書 (益岡・田窪)	1378

教科書は例文なのですべて作例

- すべてのデータを人間のアノテーターが二重にチェック

けやきツリーバンク (Butler et al. 2017)
NPCMJ 公開分に加えて、

- 青空文庫 約 5200 文
- 毎日新聞 1995 年 約 1600 文 (京大コーパスの一部と同じ)
- BCCWJ のテキストの一部 約 4400 文

などを含んだ約 4 万文。(内 1.4 万文は復元に元テキストが必要。)

- NPCMJ 公式版と比べると精度にばらつきがある
- web インターフェイスは付属しない

NPCMJ コーパスの内訳

NPCMJ 公式版 (昨年度公開計 1 万文)

テキストジャンル	ツリー数
新聞記事 (河北新報)	4323
Wikipedia 記事	2745
新・旧約聖書	1652
教科書 (益岡・田窪)	1378

教科書は例文
なのですべて作例

- すべてのデータを人間のアノテーターが二重にチェック

けやきツリーバンク (Butler et al. 2017)

NPCMJ 公開分に加えて、

- 青空文庫 約 5200 文
- 毎日新聞 1995 年 約 1600 文 (京大コーパスの一部と同じ)
- BCCWJ のテキストの一部 約 4400 文

などを含んだ約 4 万文。(内 1.4 万文は復元に元テキストが必要。)

- NPCMJ 公式版と比べると精度にばらつきがある
- web インターフェイスは付属しない

今までのコーパスとどう違うか?

BCCWJ vs. NPCMJ/けやき

	BCCWJ	NPCMJ/けやき
サイズ	大 (１億語)	小 (64 万語 *)
粒度	形態論情報のみ	統語構造
主な用途	量的分析	質的分析

* けやきコーパスの語数

- 両者は補完しあうリソースであり、競合するものではない
- 統語論研究でも、場合によっては BCCWJ や、さらには Google 検索などのほうが有用なケースもある
- 異なるリソースの特徴を知って使いわけることが重要

事例研究のアウトライン

学術誌『日本語文法』からいくつか論文を選び、けやきツリーバンクを用いて

- 論文で扱われている現象の実例をコーパスから採取できるか
- 論文の著者が主張している一般化を検索によって裏付けたり反証したりすることができるか

の二点を検討した。(主に、「コーパス検証型研究」(石川 2012) の観点からの検証を試みた。)

事例研究のアウトライン

事例研究例 2 件

- **事例研究** 1: **NPI の分布 (松井論文)**
- **事例研究** 2: **トキ節の時制解釈 (舩橋論文)**

主な結論:

- コーパスから網羅的に例を抽出することで、内省によるデータの検討からだけでは見えてこないようなパターンや要因が浮かび上がる
- 検索の仕方によっては、問題となっている現象と関連する新たな現象を発掘するといった、「コーパス**駆動型**研究」的な使い方もできそう

事例研究のアウトライン

事例研究例 2 件

- **事例研究 1: NPI の分布 (松井論文)**
- **事例研究 2: トキ節の時制解釈 (舩橋論文)**

主な結論:

- **コーパスから網羅的に例を抽出することで、内省によるデータの検討からだけでは見えてこないようなパターンや要因が浮かび上がる**
- **検索の仕方によっては、問題となっている現象と関連する新たな現象を発掘するといった、「コーパス駆動型研究」的な使い方もできそう**

謝辞

本研究は以下の研究費の助成を受けました。

- **科研費 基盤研究** (B) 15H03210 **「統語・意味解析情報タグ付きコーパス開発用アノテーション研究：複文を中心に」** (**平成** 27 **年度～平成** 31 **年度**)

NPCMJ **コーパスの開発、また本研究は、以下の方々の貢献に支えられています** (**敬称略**)**。記して感謝いたします。**

- NPCMJ **コーパス開発チーム**
  Alastair Butler、Stephen Wright Horn、**長崎郁**、Prashant Pardeshi、**吉本啓、窪田愛** (**昨年度**)
- NPCMJ **コーパス・アノテーター@国語研、東北大学**
- NPCMJ **コーパス・フィードバックチーム@神戸大学**

Butler, Alastair, Kei Yoshimoto, Shota Hiyama, Stephen Wright Horn, Iku Nagasaki, and Ai Kubota (2017) "The Keyaki Treebank Parsed Corpus, Version 1.0," http://www.compling.jp/Keyaki/, accessed 2017/07/11.

小川芳樹・長野明子・菊地朗 (2016) **「概観: 言語変化・変異の研究とコーパス」，小川芳樹・長野明子・菊地朗（編）『コーパスからわかる言語変化・変異と言語理論』，開拓社，東京，**1–28 **頁．**

石川慎一郎 (2012) **『ベーシックコーパス言語学』，ひつじ書房，東京．**