



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

Japanese

English

トップ

研究成果

リンク

お問い合わせ

お知らせ

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

「イントロダクション」

プラシヤント・パルデシ
国立国語研究所

コーパスを使った様々な日本語研究

- 南(1991) 様々な連用節内に生じる要素(主題、補足語、修飾語、助動詞)の出現頻度
- ナロック(2006) モダリティと接続助詞の組み合わせの分布
- 大曾(2007) 「～を／に参拝する」「～が／を好き」などの格助詞の出現頻度
- スルダノヴィッチ他(2008) 推量副詞のレジスター別出現頻度
- 茂木(2008) 「～(ない)ために」の二つの用法(理由、目的)の分布
- 小西(2009) カラ節・ノデ節における丁寧体のレジスター別出現頻度
- 杉本(2009) 非規範的形容詞(例:違かった)の用法
- 田野村(2009) 「そうも言う」「そうとも言う」の用法の違い
- 野口・仁科(2009) ガ格と共起する名詞の種類
- 萩野(2009) 格助詞のレジスター別出現頻度
- 李他(2009) 形容詞の連体形／連用形で意味が変わるもの
- 建石(2011) 「～たばかりに／ばかりか」「～たところが／ところで」の用法
- 丸山(2011) 連用節が多重に連鎖する現象
- 李(2011) 「XがYにVする」の意味の多様性
- 丸山(2012) 様々な連用節のレジスター別出現頻度及び連用節内のモダリティ形式

既存のコーパス：現状

- ☞ 主に形態論的な情報を付与（アノテーション）されている
- ☞ そのため、線条的な関係（共起関係）検索・検出は可能

線条的検索・検出

あげる 頻度=12,107

グループ別 パターン頻度順 基本

名詞+助詞

パターン	頻度	比率
...があげる	1,611	
...はあげる	816	
...もあげる	281	
...のあげる	59	
...をあげる	8,229	
...にあげる	2,174	
...へあげる	36	
...であげる	569	
...とあげる	62	
...からあげる	165	
...まであげる	67	
...よりあげる	14	

名詞+複合助詞

パターン	頻度	比率
...としてあげる	149	
...に応じてあげる	1	
...に応じてあげる	3	
...についてあげる	2	
...によってあげる	1	
...を通してあげる	1	
...からしてあげる	1	
...とはあげる	1	

名詞

パターン	頻度	比率
あげる+名詞	1,044	
あげ+名詞	149	
あげた+名詞	756	
あげている+名詞	114	
あげていた+名詞	28	

助動詞

パターン	頻度	比率
あげれる	1,401	

...をあげる 1209種類

コロケーション	頻度	MI	LD
声をあげる	1,116	8.85	9.82
悲鳴をあげる	427	11.89	10.49
顔をあげる	403	7.18	8.20
例をあげる	311	7.87	8.62
手をあげる	239	5.99	7.08
成果をあげる	226	9.32	9.16
効果をあげる	174	6.99	7.76
大声をあげる	137	9.91	8.79
歓声をあげる	115	11.48	8.75
力をあげる	111	5.03	6.09
腰をあげる	106	7.65	7.83
利益をあげる	105	7.30	7.66
笑い声をあげる	102	10.91	8.55
叫び声をあげる	100	11.29	8.55
名をあげる	94	5.54	6.46
目をあげる	86	4.14	5.27
ことをあげる	81	0.61	1.90
ものをあげる	80	2.17	3.43
うめき声をあげる	77	11.66	8.21
点をあげる	69	4.51	5.55
名前をあげる	64	5.79	6.48
成績をあげる	61	8.08	7.48
綱をあげる	57	8.33	7.46
叫びをあげる	56	10.23	7.70
首をあげる	50	5.11	5.90
頭をあげる	50	4.61	5.53
唸りをあげる	49	10.52	7.53
実績をあげる	45	7.75	7.07
業績をあげる	45	8.25	7.18
スピードをあげる	44	7.39	6.94
名乗りをあげる	41	11.65	7.32
片手をあげる	41	8.48	7.11
【一般】をあげる	37	-0.51	0.78
プレゼントをあげる	37	7.70	6.84
両手をあげる	37	6.97	6.64
水をあげる	35	3.92	4.88
何をあげる	33	1.98	3.18
国をあげる	33	3.31	4.37
収益をあげる	31	7.18	6.52

声をあげる 全1116件

論文: 30.11 図・書籍: 17.78 ベスト: 15.74
 出・書籍: 14.63 雑誌: 6.92 広報: 0.97
 教科書: 0.96 ブログ: 0.92 知恵袋: 0.62

- と声をあげる。(峰隆一郎著『円四郎斬鬼剣』, 1995, 913)
- と声をあげる。(峰隆一郎著『円四郎斬鬼剣』, 1995, 913)
- と声をあげた。(峰隆一郎著『円四郎斬鬼剣』, 1995, 913)
- と声をあげた。(峰隆一郎著『円四郎斬鬼剣』, 1995, 913)
- と声をあげる。(峰隆一郎著『特急「富士」はやぶさ殺人交差』, 1997, 913)
- と声をあげる。(峰隆一郎著『土方歳三』, 2000, 913)
- と声をあげる。(峰隆一郎著『剣鬼・針ヶ谷夕雲』, 2001, 913)
- と声をあげた。(峰隆一郎著『剣鬼・針ヶ谷夕雲』, 2001, 913)
- と声をあげる。(峰隆一郎著『剣鬼・針ヶ谷夕雲』, 2001, 913)
- と声をあげる。(峰隆一郎著『剣鬼・針ヶ谷夕雲』, 2001, 913)
- と声をあげる。(峰隆一郎著『剣鬼・針ヶ谷夕雲』, 2001, 913)
- と声をあげる。(峰隆一郎著『剣鬼・針ヶ谷夕雲』, 2001, 913)
- と声をあげた。(富山光三郎著『江戸前寿司一の一の店を行く』, 2002, 596)
- と声をあげた。(桐生操著『本当は恐ろしいグリム童話』, 1998, 913)
- と声をあげる。(峰隆一郎著『富札を斬る』, 2001, 913)
- と声をあげる。(峰隆一郎著『富札を斬る』, 2001, 913)

NINJAL-LWP for BCCWJ Copyright © 2012-2014 National Institute for Japanese Language and Linguistics, Lingo Institute of Language. All rights reserved.

<http://nlb.ninjal.ac.jp/>

既存のコーパスの問題点

“日本語コーパスを用いた複文構文の研究は、現在までのところ、連用節の接続形式が持つ**形態的な側面**に着目した研究が多いように思われる。一方、例えば、連用節の接続形式が主節のモダリティ形式に制限を与えるといった**文法的制約をコーパスから検索し、定量的に分析するような研究の事例は、管見の限りない**。これは、**離れた位置にある構文要素の対応関係**を自動的に取得するための研究用情報(**統語構造情報**)が、一般に使いやすい形で整備されていないことが理由として考えられる。”(丸山 2014; p. 391)

- キーワード1: **離れた位置にある構文要素の対応関係**
- キーワード2: **統語構造情報**

既存のコーパスの問題点

- ☞ 統語論的な情報が付与されていない→文の中核的な文法情報である語句間の依存関係、句、節、複文といった様々なレベルでの構造体を検索・抽出は困難
- ☞ コーパスに基づく日本語の諸構造体の研究は英語などと比べて、立ち遅れている
- ☞ 日本語と世界の他言語とのコーパスにもとづく対照言語研究を行うためにも言語研究を目的とする汎用的なツリーバンクの開発は不可欠

線条的検索・検出から階層的検索・検出へ

- 👉 目標: コーパスに基づく言語研究を目的とする汎用的な統語解析情報をタグ付けしたコーパスを開発し、様々な検索ツールとともに公開する
- 👉 国立国語研究所で2016年4月から文統語解析情報に加えて意味解析情報をタグ付けした日本語コーパスNINJAL Parsed Corpus for Modern Japanese (略称NPCMJ) の構築を開始。

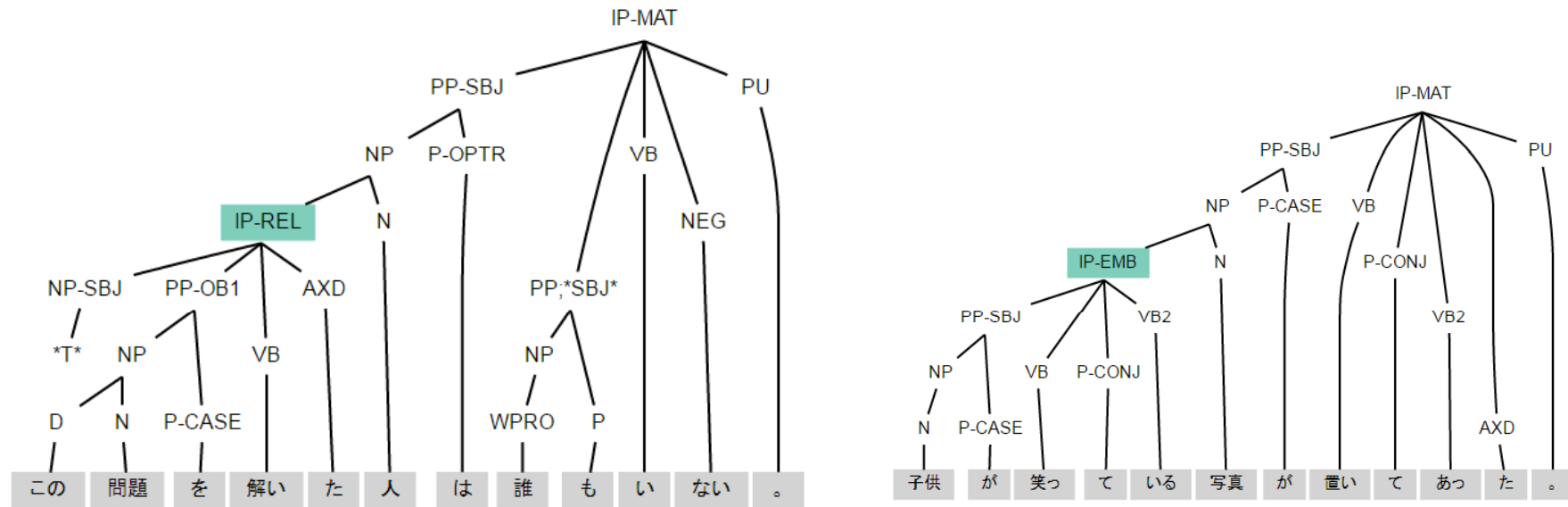
アノテーションの原則

ペン通時コーパス方式

- ☞ ペン通時コーパスの方式を採用(詳しくは吉本の発表で) - Annotation Manual for the Penn Historical Corpora and the PCEEC (Santorini 2010)
- ☞ 利点1: 世界の多様な言語のコーパスに利用されており(例: 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)、他の言語のコーパスとの比較・対照が容易
- ☞ 利点2: **句や節に機能タグ**が付けられ、より詳細な統語情報及び意味情報が得られる
例: NP-SBJ, NP-OB1, NP-TMP...、IP-REL, IP-EMB...

なぜ統語・意味解析コーパスか

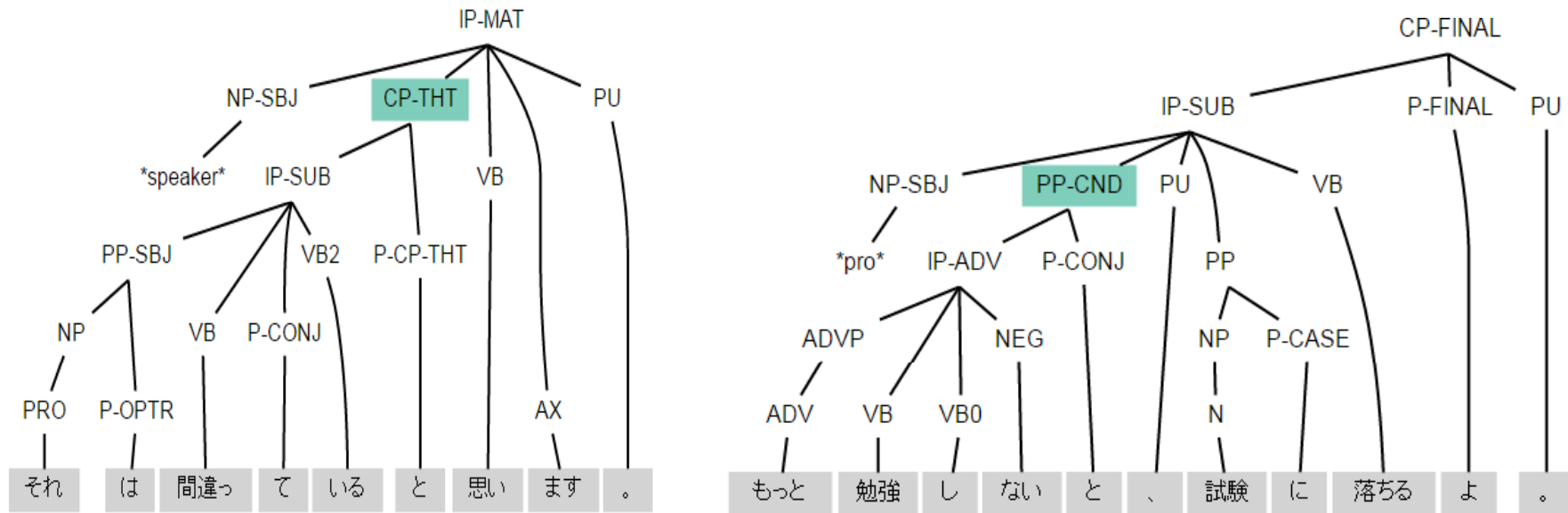
☞ 既存のコーパス: 「内の関係」と「外の関係」はともに[動詞+ 助動詞] [名詞] のように分析され、両者の違いを捉えることは出来ない。



NPCM]における「内の関係」と「外の関係」の区別用の
アノテーション方法

統語・意味解析コーパスか

☞ 既存のコーパス: 接続助詞「と」が導く節には引用節と条件節[動詞句+助詞]のように分析するだけで、意味上の区別には役立たない。



NPCMJにおける助詞「と」の曖昧性解消のアノテーション方法

おわりに

- 👉 意味解析のための適合性、記述の客観性と一貫性、言語使用の実態の反映および検索利用の容易性、という異なる要因のバランスを配慮した開発を行う。
- 👉 言語データとしては、著作権上の問題の生じない Wikipedia 記事や問題が解決された新聞記事等を取り上げる
- 👉 6年間で約5～6万文の日本語テキストについて、統語・意味解析情報付きコーパスを完成させる。

2016年度公開のデータ

出典	例文数	語数
『河北新報』記事	4243	69461
Wikipedia 記事	2752	63646
新約・旧約聖書	1659	26715
益岡・田窪(1992)例文	1378	11910
	合計: 10032	171732

👉 本日デモを行う→<http://npcmj.ninjal.ac.jp/interfaces/>
「npcmj」でGoogle検索→「NPCMJコーパスで調べる」

ウェブインターフェースの紹介



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

現在開発途中のウェブインターフェース:

Explore NPCMJ

<http://npcmj.ninjal.ac.jp/interfaces/>

「npcmj」でGoogle検索

→ 「NPCMJコーパスで調べる」



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



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

Japanese English

トップ 研究成果 リンク お問い合わせ お知らせ

お問い合わせ

ご感想、ご意見をお寄せください。以下のフォームに記入してください。

氏名：*

姓 名

メールアドレス：

メッセージ：

文字認証：

9fu5kx

送信



参考文献

石川慎一郎(2012)『ベーシックコーパス言語学』ひつじ書房

Beatrice Santorini (2010) Annotation Manual for the Penn Historical Corpora and the PCEEC (Release 2). University of Pennsylvania.

丸山岳彦(2014)「コーパス言語学・語用論の観点から見た日本語複文研究 の動向と課題」『日本語複文構文の研究』pp.385 -398. ひつじ書房

コメント、フィードバックを
プロジェクトウェブサイト
の「お問い合わせ」で
よろしくお願
いいたします。

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