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

「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

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うる要素（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

（石川 2012、丸山 2014）

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

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

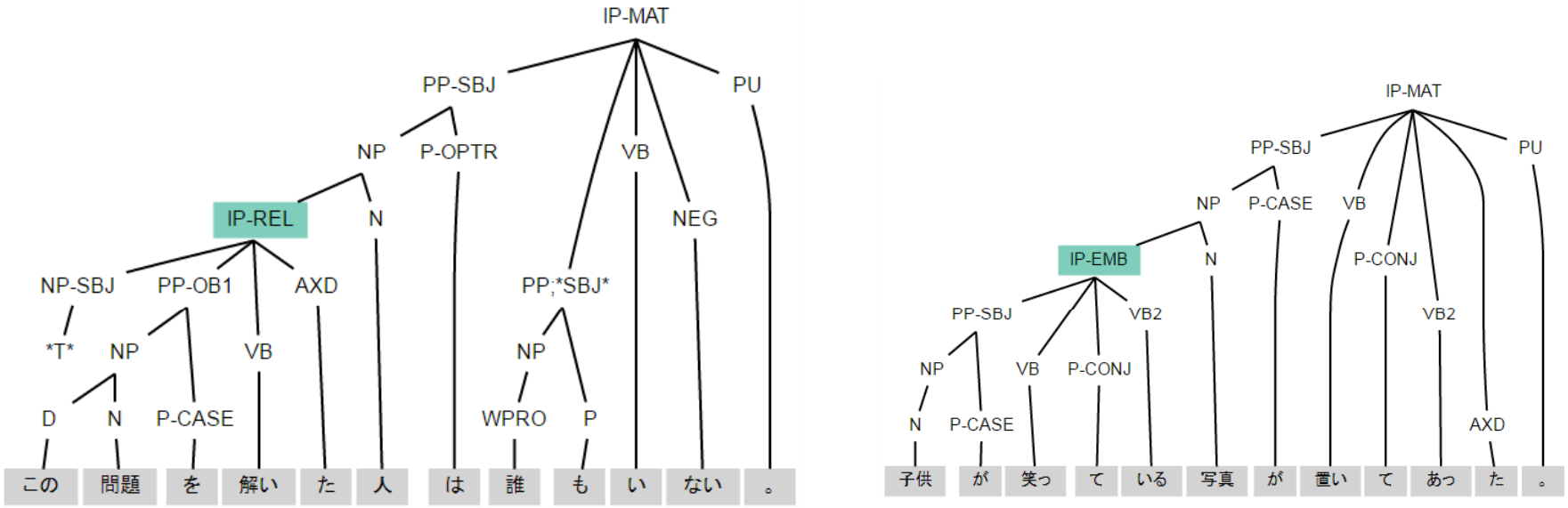
- ☞ペン通時コーパスの方式を採用（詳しくは吉本の発表で）- Annotation Manual for the Penn Historical Corpora and the PCEEC (Santorini 2010)

- ☞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

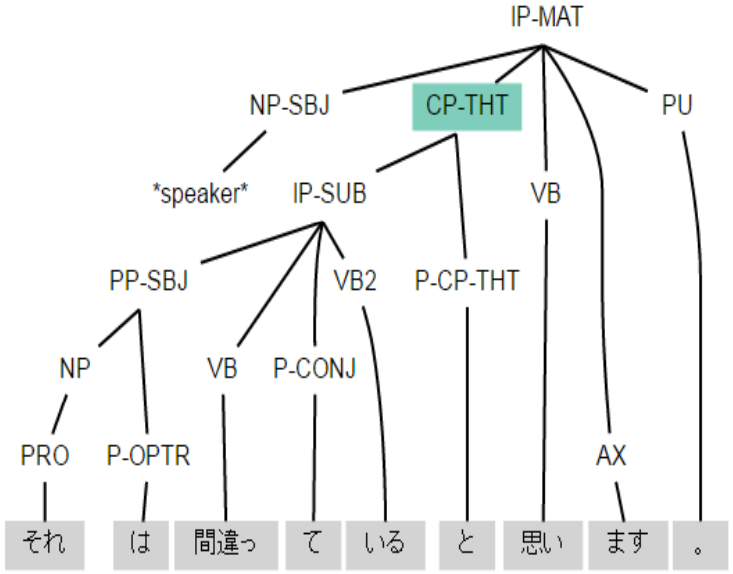
☞既存のコーパス:「内の関係」と「外の関係」はともに[動詞+ 助動詞] [名詞] のように分析され、両者の違いを捉えることは出来ない。

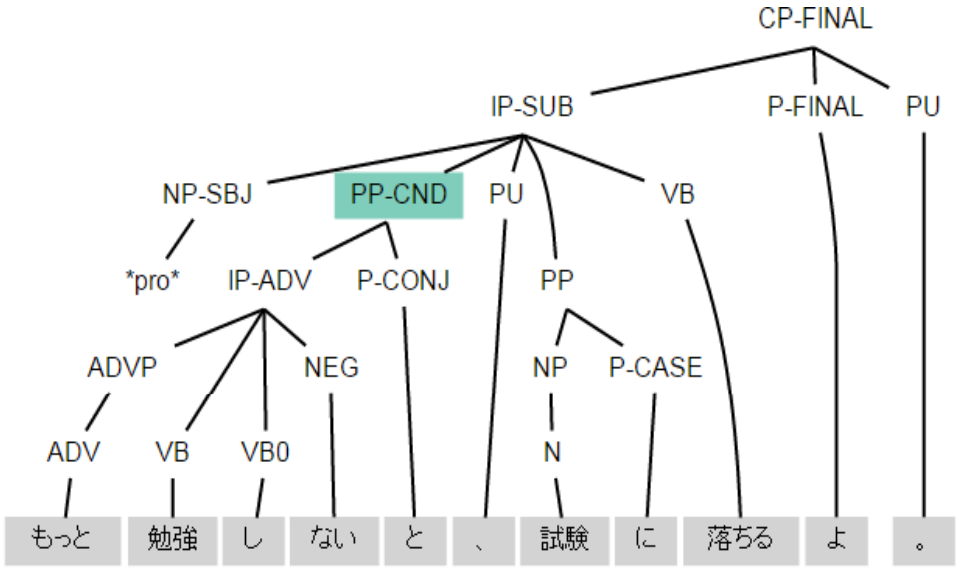


NPCMJにおける「内の関係」と「外の関係」の区別用のアノテーション方法

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 万文の日本語テクストについて、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/

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 Mannual 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/