



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/>

「統語・意味解析情報付き日本語 コーパスの構築に向けて」

企画：プラシヤント・パルデシ

司会：吉本 啓

コメンテーター：福島 一彦

趣旨説明

👉 ワークショップの目的:

- ①「統語・意味解析情報付き日本語コーパス」の
開発の動機、特色、アノテーション方式、意義
について解説する
- ②具体例のデモンストレーションを行い、本コーパス
について理解を深める
- ③皆さんからのフィードバックを得て今後役に立てる

趣旨説明

- 👉 「統語・意味解析情報付き日本語コーパス」とは
- 👉 国立国語研究所共同研究プロジェクト「**統語・意味解析コーパスの開発と言語研究**」で開発中のコーパス (<http://npcmj.ninjal.ac.jp/>)
- 👉 **NINJAL Parsed Corpus of Modern Japanese (NPCMJ)**
- 👉 2016年4月～2022年3月の6年間で、毎年1万文（合計6万文）を公開予定。

趣旨説明



トップ

研究組織図

メンバー

News

- ヨーク大学言語学科と交流協定を締結しました
2016年7月27日
- ペンシルバニア大学言語学科と交流協定を結びました
2016年7月20日

このプロジェクトについて

現在世界の主要言語について統語解析情報付きコーパス (ツリーバンク) が作られ、それにより言語学および言語処理の研究において目覚ましい成果が現れています。しかし日本語については公開された十分な規模のそのようなコーパスは存在しません。このような日本語研究の遅れを挽回し、多様な日本語の機能語や句構造、節の諸類型および複雑な構文を大量の言語データから検索・抽出して研究することを可能にすることを目的として、現代日本語の書き言葉と話し言葉のテキストに対し文の統語・意味解析情報をアノテートした **NINJAL Parsed Corpus of Modern Japanese (NPCMJ) / けやきツリーバンク** の構築を開始しました。

汎用性を高めるため、情報の抽出を最優先させたアノテーションのスキーマを採用しています。具体的には、Penn Treebankをもとに発展したペン通時コーパス (Penn Historical Corpus; Santorini 2010) のアノテーション方針に従っています。

成果として得られるコーパスは、技術的スキルなしに誰でも簡単に利用できるインタフェースとともに、国立国語研究所のホームページから一般公開する予定です。

プロジェクトリーダー



ブラシャント・バルデシ (国立国語研究所 理論・対照研究領域 教授)

「**npcmj**」で
Google検索

Copyright © 2016 National Institute for Japanese Language and Linguistics. All rights reserved.

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

ワークショップの構成

- (1) 趣旨説明＋イントロダクション(15分)
プラシヤント・パルデシ(国立国語研究所)
- (2) アノテーション方式とコーパスの特色(25分)
吉本啓(東北大学)  **バックエンド(仕組み)**
- (3) デモンストレーション(45分)
アラステア・バトラー、窪田愛(国立国語研究所)、
窪田悠介(筑波大学)  **フロントエンド(検索)**
- (4) まとめと将来の展望(5分): プラシヤント・パルデシ
(国立国語研究所)
- (5) コメント: 福島一彦(関西外国語大学)(10分)
- (6) 質疑応答(20分)