国立国語研究所「統語・意味解析コーパスの開発と言語研究」プロジェクト
「統語・意味解析コーパス (NPCMJ)」チュートリアル（オンライン開催）
2022年3月4日

NINJAL PARSED CORPUS OF MODERN JAPANESE（NPCMJ）の理念と概要 および 初心者向け検索ツール

プラシャント・パルデシ、鈴木彩香（国立国語研究所）

1

「統語・意味解析コーパス（NPCMJ）チュートリアル」

2

- 「統語・意味解析コーパス (NPCMJ) チュートリアル」に参加していただき、ありがとうございます。
- チュートリアルを始める前に、注意事項をお伝えします。
  - ①氏名のフルネーム表記にご協力をお願いいたします。
  - ②表示名の登録をせずに入室できた場合で、表示名が登録時のお名前と異なる場合には、「参加者」、ご自身の表示名の横の「詳細」、「名前の変更」の順にクリックして、お名前を入力してください。
  - ③質疑応答の時間を除いて、マイクはミュートにしてご参加ください。
  - ④主催者が不審なアクセスと判断した場合は、強制的に退出していただくことがあります。

2

「統語・意味解析コーパス（NPCMJ）チュートリアル」

3

- 「統語・意味解析コーパス (NPCMJ) チュートリアル」に参加していただき、ありがとうございます。
- チュートリアルを始める前に、注意事項をお伝えします。
  - ⑤本チュートリアルの内容は録画されるが、公開はしません。以前開催した講習会は国立国語研究所の公式YouTubeチャンネルで公開されています。
    https://www.youtube.com/playlist?list=PLZfZgVvFbh1ZLsndcOVYaS-z3GFkH5Any
  - ⑥本チュートリアルで配信する音声・映像・スライド画像等を、録音・録画・画面キャプチャなどの方法で電子的に保存する行為、およびそれらのデータを再配布する行為を禁止いたします。

3

「統語・意味解析コーパス（NPCMJ）チュートリアル」

4

- 本日のチュートリアルでは、
  - NPCMJコーパス開発の理念と概要
  - コーパスの鳥瞰図（初心者向けの検索ツールを通して）
  - どのような情報（タグ）がどのような方法で付与されるか（アノテーション方式の概要）
  - どのように検索することができるのか（様々な検索ツール、検索式の書き方）等

  を中心に、できるだけわかりやすく、解説する予定です。
- 約7時間の長丁場となりますが、よろしくお願いします。

4

本日のプログラム

5

(1) NPCMJ コーパスの理念とデータの概要、および初心者向けインターフェース 10:00-11:00 (60分)
- プラシャント・パルデシ、鈴木彩香
- 休憩：11:00-11:10

(2) タグおよびアノテーションの概要 11:10-12:10 (60分)
- 吉本啓
- 休憩：12:10-13:15

(3) 検索インターフェース（NPCMJ Development Interfaces）13:15-14:00 (45分)
- Alastair Butler, 長崎郁
- 休憩：14:00-14:10

(4) Tregex 検索式1（tag の記述）14 :10-15:10 (60分)
- 金城由美子 休憩：15:10-15:20

(5) Tregex 検索式2（tree の記述）15:20-16:50（90分）
- 長崎郁

(6) 閉会の辞
- プラシャント・パルデシ 16:50-17:00 (10分)

5

セッション①

6

NPCMJ コーパスの理念とデータの概要、および初心者向けインターフェース

10:00-11:00 (60分)

- プラシャント・パルデシ、鈴木彩香

6

セッション①の構成

7

- （A）NINJAL Parsed Corpus of Modern Japanese （NPCMJ）の理念と概要：パルデシ
- https://npcmj.ninjal.ac.jp/

- （B）初心者向け検索ツール（NPCMJ Explorer）：鈴木
- https://npcmj.ninjal.ac.jp/explorer/

7

セッション①の構成

8

- （A）NINJAL Parsed Corpus of Modern Japanese （NPCMJ）の理念と概要：パルデシ
- https://npcmj.ninjal.ac.jp/

- （B）初心者向け検索ツール（NPCMJ Explorer）：鈴木
- https://npcmj.ninjal.ac.jp/explorer/

8

本題に入る前に：そもそもコーパスとは？

9

- 言語学において統計的な分析や研究を行う目的で集められ構築された、言語テキストの集合体を指す。ラテン語で「身体」を意味する ‘corpus’ が由来。近年では電子化されデータ利用できるものがほとんどで、「電子コーパス」と同義でとらえられる。実際の書き言葉や話し言葉を言語資料として大量に集積し、それを検索して得られた結果を証拠とした言語記述を可能にしたことから、経験主義的な言語研究の発達に大きく貢献している。

  ebookpedia http://www.jepa.or.jp/ebookpedia/201612_3301/

- 言語を分析するための基礎資料として，書き言葉や話し言葉の資料を体系的に収集し，研究用の情報を付与したもの。

  国立国語研究所HP https://www.ninjal.ac.jp/database/type/corpora/

9

ネットにある膨大なデータはコーパスとは言えない

10

電子化された大量の基礎資料であるが

- 研究用の情報（アノテーション）が付与されていない
- 文字列検索のみが可能
- 言語学的な研究には十分とはいえない
- ほしい情報が検索できるためには

☞情報を付加する作業（アノテーション）が不可欠

10

望ましいコーパスとは？

11

- 電子化された大量の基礎資料である
- 言語分析・統計分析のための情報（アノテーション）が付与されている
- 検索を可能とする専用検索ツールが提供されている
- 無償で一般公開され、研究目的の場合、検索結果や全データをダウンロードすることが可能である

11

言語分析のための基礎資料：多様性

12

- 古代語
- 近代語
- 現代語：書き言葉、話し言葉
- 方言
- こどもの母語習得のデータ
- 学習者の外国語習得のデータ
- 複数の言語のデータ（パラレルコーパス）

12

言語分析のための基礎資料@国語研

13

- 古代語：日本語歴史コーパス（CHJ）、THE OXFORD-NINJAL CORPUS OF OLD JAPANESE (ONCOJ)
- 近代語：近代語のコーパス（『太陽コーパス』『近代女性雑誌コーパス』『明六雑誌コーパス』『国民之友コーパス』）
- 現代語：現代日本語書き言葉均衡コーパス（BCCWJ）、国語研日本語ウェブコーパス、日本語話し言葉コーパス（CSJ）
- 言語習得：多言語母語の日本語学習者横断コーパス（I-JAS）
- 方言：日本語諸方言コーパス (COJADS)
- 同一内容の複数言語のデータ（パラレルコーパス）：なし

13

研究用の情報（アノテーション）の多様性

14

- 音声・音韻論的な情報を付与
- 形態論的な情報を付与
- 統語論的な情報を付与
- 意味論的な情報を付与
- 語用論的な情報を付与　など

14

現代日本語のコーパス：公開状況と規模

15

- 現代日本語書き言葉均衡コーパス（BCCWJ、約1億語）
- 京都大学テキストコーパス（約4万文）
- 『筑波ウェブコーパス』(TWC、約11億語）
- 日本語係り受けコーパス (JDC、約4万3000文、95万語)
- 国語研日本語ウェブコーパス（NWJC）（約100億語）
- UD Japanese-BCCWJ （ 約 127万単語）

※ アノテーション構造のみを公開。データは未公開

- 統語・意味解析情報付き現代日本語コーパス（NPCMJ）

2022年2月現在：ツリー数： 98,312（約145万語）

アノテーションとデータの両方を複数の検索ツールとともに無償公開。

15

既存の他のコーパスとNPCMJの主な違い

16

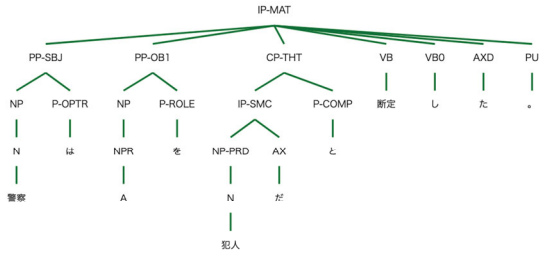
- 言葉を文字で書きあらわそうとすると、文字が一直線に並ぶが実際には、言語は直線ではなく階層的な構造を作っている
- 既存のコーパスの多くは文を「分節」に区切り、形態論的な情報および線状的な関係（係り受け）の情報を付与するものが主流
- NPCMJは形態論的な情報に加えて、文の階層的な構造、つまり、統語構造の情報を付加したコーパス

16

既存の他のコーパスとNPCMJの主な違い

17

NPCMJは形態論的な情報に加えて、文の階層的な構造、つまり、統語構造の情報を付加したコーパス



句構造を利用した文の構造の表示

17

NINJAL Parsed Corpus of Modern Japanese (NPCMJ) とは

18

- 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（＝統語構造）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クストに対し文の統語・意味解析情報をアノテートしたコーパス

☞ https://npcmj.ninjal.ac.jp/

18

NINJAL Parsed Corpus of Modern Japanese (NPCMJ) とは

19

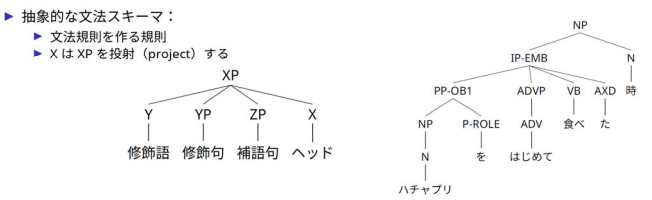
- 国立国語研究所共同研究プロジェクト「統語・意味コーパスの開発と言語研究」で2016年4月1日から開発を開始
- 2022年2月現在：2022年2月現在：ツリー数： 98,312（約145万語）をデータとともに無償公開：https://npcmj.ninjal.ac.jp/
- 現代日本語以外に古代語、方言（津軽方言）、L1およびL2としての日本語獲得・習得のデータにNPCMJ方式のアノテーションを付与し、上記HPで公開（@NPCMJ Development Interfaces）

19

NINJAL Parsed Corpus of Modern Japanese (NPCMJ) とは

20

- アノテーション方式：汎用性を高めるため、情報の抽出を最優先させるペン通時コーパス (Penn Historical Corpus; Santorini 2010) のアノテーション方針を採用☞詳細は吉本講師から
- 句構造(phrase structure):Xバー理論



20

NINJAL Parsed Corpus of Modern Japanese (NPCMJ) とは

21

- 現在公開中の試行版コーパスは、技術的スキルなしに誰でも簡単に利用できるブラウザから高度な検索ができる複数のインターフェース（検索ツール）とともに無償公開、検索結果のダウンロードも可 ☞ https://npcmj.ninjal.ac.jp/

- 今日のチュートリアルの内容は以下のとおり

NPCMJ Explorer（初心者向けインターフェース）:鈴木
タグおよびアノテーションの概要：吉本
検索インターフェース（NPCMJ Development Interfaces）：Butler, 長崎
Tregex 検索式1 (tag の記述)：金城
Tregex 検索式2 (tree の記述)：長崎 郁

その前に、コーパスの使用上の注意点をお話しします。

21

コーパス使用上の注意点

22

- 研究者は自分の研究目的に合ったコーパスを選定する必要がある
- 現代日本語書き言葉均衡コーパス（BCCWJ）：形態論的な情報を付与したコーパス、線上的に隣接する形態素・語を検索・抽出できることが強み
- NPCMJコーパス：統語・意味解析情報（階層構造に基づく文法機能の情報）を付与したコーパス

22

コーパス使用上の注意点

23

①必要なデータを検索・抽出するためにアノテーションの方法を理解する必要がある。特定の言語現象（例えば、名詞修飾表現）に関する理論研究での分析とコーパス開発で採用されている分析は同じであるとは限らない ☞ NPCMJ アノテーションマニュアル(本日現在342頁)を参照
https://npcmj.ninjal.ac.jp/wp-content/uploads/2020/05/npcmj_annotation_manual_jp_20200501.pdf

②検索結果の中にはゴミ、アノテーションミスなども含まれている。検索結果をダウンロードして自分の目で確認する必要がある

③コーパスは大量のデータから興味のある現象に関するデータを抽出するための資源であり、分析を提供するものではない

23

セッション①の構成

24

- （A）NINJAL Parsed Corpus of Modern Japanese （NPCMJ）の理念と概要：パルデシ
- https://npcmj.ninjal.ac.jp/

- （B）初心者向け検索ツール（NPCMJ Explorer）：鈴木
- https://npcmj.ninjal.ac.jp/explorer/

24