2019年5月11日(10:00~17:30)
@弘前大学創立50周年記念会館 会議室2

大学共同利用機関法人 人間文化研究機構
国立国語研究所
National Institute for Japanese Language and Linguistics
NINJAL

NPCMJ コーパスの理念と概要、および初心者向けインターフェース

プラシャント・パルデシ(国立国語研究所)
Prashant Pardeshi (NINJAL, Tokyo)

「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」@弘前大学に参加していただき、ありがとうございます。
- 本日のチュートリアルでは、
  - そもそも、コーパスとはなにか
  - NPCMJコーパス開発の理念
  - アノテーションの方式、どのようなデータを利用して、どのように開発されているのか
  - どのように検索することができるのか

  などについてそれぞれの講師がわかりやすく解説する予定です。
- よろしくお願いします。

コーパスとは

3

- 語彙索引など、言語研究のための資料。特に、コンピューターを利用してデータベース化された大規模な言語資料。

  デジタル大辞泉

  https://kotobank.jp/word/コーパス-498318
- コーパスとは「（特定の種類・作家の文書[資料]の）集大成、集積」をさす。現在では、大量に収集したテキストデータをコンピュータで解析可能なかたちにした、いわゆるコンピュータコーパスをさすことが多い。

  ウィズダム英和辞典 第3版 - 三省堂WORD-WISE WEB

  https://dictionary.sanseido-publ.co.jp/dicts/english/wisdom_ej3/sp/corpus.html

コーパスとは

4

- 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/

望ましいコーパスとは

5

- 言語分析・統計分析のための電子化された基礎資料
- 大量のデータ
- 研究用の情報（アノテーション）を付与したもの
- 検索可能（専用検索ツール・インターフェースの開発・公開）
- 無償公開、検索結果・データダウンロード可

代表性と均衡性

6

- 前川（2013：13）：代表性（representativeness）

「現代日本語のように膨大な言語資料から構成されている対象の場合は、全体を有限のサンプルで代表させざるをえない。その種のコーパス（サンプルコーパス）は対象変種全体の縮図となっていることが望まれる。」

- 前川（2013：13）：均衡性（balance）

「自然言語には通常多数の変種が存在する。。。積極的に多数の変種をカバーして、対象言語の全体像を把握しようとするコーパスは均衡コーパス（balanced corpus）と呼ばれる。」

言語分析のための基礎資料：多様性

7

- 古代語（歴史コーパス）
- 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）

言語分析のための基礎資料@NINJAL

8

- 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）
- 学習者：多言語母語の日本語学習者横断コーパス（I-JAS）
- 方言：現在開発中@国立国語研究所
- 複数の言語のデータ（パラレルコーパス）：なし

現代日本語コーパスの規模

9

- 現代日本語書き言葉均衡コーパス（BCCWJ、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、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（100億語）
- 統語・意味解析情報付き現代日本語コーパス（NPCMJ）
（2018年3月現在：ツリー数：約2万、約30万語）
- Universal Dependencies of Japanese（1万文）
- 超巨大コーパス（基盤研究(A)18H03575「準均衡超大規模日本語コーパスと高速検索ツールの開発」で開発中、約200億語、2022年に公開予定）

研究用の情報（アノテーション）の付与

10

- 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

検索可能性

11

- ネットでの膨大のデータ：電子化されたものだが
- 均衡性を欠く
- 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☞アノテーションが不可欠

NINJAL Parsed Corpus of Modern Japanese (NPCMJ) とは

12

- 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

☞ http://npcmj.ninjal.ac.jp/

NINJAL Parsed Corpus of Modern Japanese (NPCMJ) とは

13

- 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
- 2019年3月現在：約3万文・ツリー（50万語）を無償約公開。
- 今後、3年にわたって毎年1万文（ツリー）を追加し、プロジェクト終了時点（2022年3月）までに６万文・ツリー（100万語）を公開する予定

NINJAL Parsed Corpus of Modern Japanese (NPCMJ) as of March 2019

14

出典	ツリー数	語数
青空文庫（aozora）	4,646	101,537
聖書（bible）	1,664	30,657
書籍（book）	552	12,515
辞書（dict）	3,419	33,651
国会会議録（diet）	1,698	37,349
フィクション（fiction）	923	12,051
法律文（law）	337	7,793
その他（misc）	2085	23,872
ニュース（news）	4,666	84,927
ノンフィクション（nonfiction）	223	4,454
テッドトーク（ted）	1,453	22,030
教科書（textbook）	6,048	64,038
ウィキペディア（wikipedia）	2746	70,445
合計	30,460	505,319

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

NINJAL Parsed Corpus of Modern Japanese (NPCMJ) とは

15

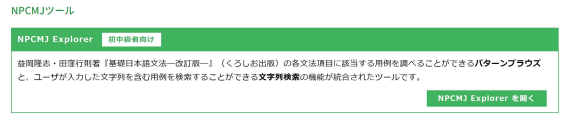
- アノテーション方式：汎用性を高めるため、情報の抽出を最優先させるペン通時コーパス (Penn Historical Corpus; Santorini 2010) のアノテーション方針を採用
  ☞詳細は吉本先生から
- 現在公開中の試行版コーパスは、技術的スキルなしに誰でも簡単に利用できるブラウザから高度な検索ができる複数のインターフェース（検索ツール）とともに無償公開、検索結果のダウンロードも可
  ☞ http://npcmj.ninjal.ac.jp/
  ☞詳細は吉本講師、鈴木講師、長崎講師

NPCMJ利用用のインターフェース http://npcmj.ninjal.ac.jp/

16

NPCMJ利用用のインターフェース
http://npcmj.ninjal.ac.jp/

17



NPCMJ Explorer（初心者向けのツール）

利点：クリックすることだけでコーパスをブラウズすることができる

弱点：予め用意された検索項目しか検索できない

中上級者向けの検索ツールへの橋渡し的な役割も兼ねている

コーパス使用上の注意点

18

- コーパスの役割は、研究に必要なデータを大量のデータから抽出し、提供すること（分析を提供することではない）
- 研究目的に合ったコーパスを選定する
- 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

コーパス使用上の注意点

19

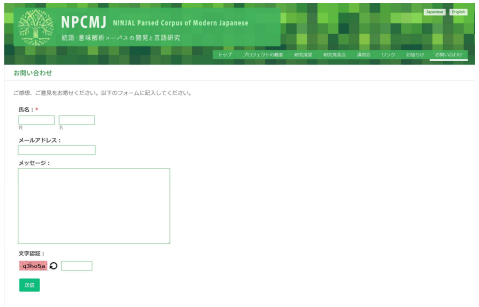
①必要なデータを検索・抽出するためにアノテーションの方法を理解する必要がある。特定の言語現象（例えば、名詞修飾表現）に関する理論研究での分析とコーパス開発で採用されている分析は同じであるとは限らない

②検索結果の中にはゴミも含まれている。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

☞続きは吉本先生から

お願い：NPCMJを使ってフィードバックください。また、宣伝もよろしく。

20



NPCMJホームページ
http://npcmj.ninjal.ac.jp/