# 連絡

- 次回は 5/13 です。
- 5/13 の授業までに、授業中に紹介したい論文について考えてきて下さい。
  - 具体的に担当する論文は、分量や難易度を考えて話し合って決めていきます。
  - 紹介する論文は期末レポートと関連していても、していなくてもかまいません。

# コーパスとしてのウェブ

4/23

# おしながき

- コーパスの選択
- ウェブ検索自体は「コーパス調査」と見なせるか
- ウェブ上で利用可能なコーパス

# はじめに

- コーパスを選択するうえで
  - 内容が目的に合っているか？
  - 分量は適切か？
  - 入手可能性（価格含め）や検索ツールの利用可能性など

# ウェブ検索は
# コーパス調査になるか?

# コーパスとしてのウェブ：
# 利点

- 手軽
- 規模が大きい
  - 日本語26兆字 (田野村 2009 の推計)
- 書き手・ジャンルが多様。最新のデータを含む

# コーパスとしてのウェブ：
# 難点

- 検索の仕組みが不明（企業秘密）。そのため、検索結果件数が信頼できるのかどうか不明。
- 正確な母集団が不明。例えば「この表現を使っている人が全体の何割」というようなデータを出すのは困難
- いつ書かれたのか、著者のバックグラウンド（ネイティブなのかどうかを含め）、などが不明なことが多い
- 再現性に乏しい

# ウェブ検索は
# コーパス調査になるか?

- 検索して出てきた面白い例を出発点にして議論のアイディアを考えるぶんには問題ない
- しかしウェブ検索結果を「証拠」として使うのは色々なハードルがある

# 検索結果件数の解釈の難しさ

- 形態素解析の問題
  - 「走れメロ」と「走れメロス」
- 検索件数の問題
  - 「見当がつかない」と「検討がつかない」
  - 無関係な語と並べて検索するというテクニック
- 重複するページ、不適切なページの自動的な排除等の問題

# おすすめ

- 荻野 (2014) ウェブ検索による日本語研究.
- 荻野・田野村 (編) (2014) コーパスとしてのウェブ.


# ウェブ上で利用可能なコーパス

# ウェブ上で利用可能なコーパス

- ウェブの発達に伴い、コーパスをそのままオンラインで検索できるサイトが増えてきた
  - パソコンのOS等に依存せず、ソフトウェアのインストールなしですぐに始められるのが利点。
  - しかし、突然サービスが終了するかもしれない、提供された機能の範囲内のことしかできない、内部の動作がしばしばブラックボックスであり開発元を信じるしかない、などの難点もある。

# 青空文庫

- 著作権切れの作品を中心に収集したウェブサイト
- 全文検索を提供しているウェブサイトがいくつかある
  - http://www.su-ki-da.com/aozora/
  - http://www.let.osaka-u.ac.jp/~tanomura/kwic/aozora/

# 国立国語研究所 (NINJAL) のコーパス

- 日本語書き言葉均衡コーパス (BCCWJ)
- 日本語話し言葉コーパス (CSJ)
- 日本語歴史コーパス (CHJ)
- その他（太陽コーパスなど）

# 日本語書き言葉均衡コーパス (BCCWJ)

- **少納言** - オンライン KWIC 検索。すぐに利用可 http://www.kotonoha.gr.jp/shonagon/
- **中納言** - 高機能なオンライン KWIC 検索。無料だが利用申請必要 https://chunagon.ninjal.ac.jp/
- **NINJAL-LWP** - レキシカルプロファイラ。すぐに利用可 http://nlb.ninjal.ac.jp/
- **DVD** - 有料。扱いに知識が必要

# 日本語話し言葉コーパス (CSJ)

- 日常会話というより講演などが中心
- 現状、オンラインで検索はできない。DVD の購入が必要。

# 日本語歴史コーパス (CHJ)

- 開発中。平安時代編と室町時代編I狂言が公開済み
- **中納言** - 高機能なオンライン KWIC 検索。無料だが利用申請必要 https://maro.ninjal.ac.jp/

# 国語研の近代語コーパス

- 太陽コーパス (有料)
- 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

# 国会会議録

- http://kokkai.ndl.go.jp/

# 学習者コーパス

- http://cblle.tufs.ac.jp/llc/ja/
- http://sakubun.jpn.org/

# リンク集

- 国語研のデータベース一覧
  - http://www.ninjal.ac.jp/database/
- コーパス日本語学のための情報館
  - http://www30.atwiki.jp/corpus-ling/pages/72.html

# ウェブ上の英語コーパス

## Sketch Engine

- 英語、日本語などのウェブコーパスをはじめ、さまざまなコーパスを搭載
- 有料。ただし30日間の試用期間あり
  - https://the.sketchengine.co.uk/

## British National Corpus

- 1995年。1億語の均衡コーパス
- http://corpus.byu.edu/bnc/

# Corpus of Contemporary American English (COCA)

- 現在拡張中のウェブ均衡コーパス
- http://corpus.byu.edu/coca/

# 古典語

- Perseus Project
  - http://www.perseus.tufts.edu/hopper/

# コメントシート

- 特に利用してみたいと思ったコーパスはどれですか。
- 今回紹介しなかったもので、このようなコーパスはないのか、という質問でもかまいません。