「日本語史研究における コーパス利用の現在地」 へのコメント

北﨑勇帆（きたざき ゆうほ）

高知大学人文社会科学部

「現在地」とその課題

- 挙げられた4つの課題
  1. アノテーションへの依存
  2. 研究対象資料がコーパス化されていないといけない
  3. コーパス化のハードル
  4. 文字・表記研究に弱い
- これが「課題」となるのは（良くも悪くも）CHJがインフラ化したことによる
  - 「まずはCHJ」にならない時代はそれほど問題にはならなかった

「アノテーションへの依存」の課題

- UIと規程の難解さ
  - 書字形出現形？　語彙素？
  - 助動詞「う」？　意志推量形？
  - 複合動詞？　接頭辞＋動詞？　動詞＋動詞？
- どう対処する？（できる？）
  - 簡易検索ツール「ことねり」：https://cotoneri.ninjal.ac.jp/
  - 『コーパスによる日本語史研究』（ひつじ書房）解説編

- 青木博史・岡﨑友子・小木曽智信編（2022）『コーパスによる日本語史研究 中古・中世編』ひつじ書房
- 田中牧郎・橋本行洋・小木曽智信編（2021）『コーパスによる日本語史研究 近代編』ひつじ書房
- 田中牧郎編（2020）『コーパスで学ぶ日本語学 日本語の歴史』朝倉書店
- 中俣尚己（2021）『「中納言」を活用したコーパス日本語研究入門』ひつじ書房

「アノテーションへの依存」の課題

- 付された形態論情報への依存と利用
  - 小木曽（2019）：研究者が新たに付すアノテーションについて
  - 北﨑（2022）：既存の形態論情報の利用可能性について

- 小木曽智信（2019）「CHJ 中古「る」「らる」用法分類アノテーションデータ」https://researchmap.jp/mu1dor8so-12361/
- 北﨑勇帆（2022）「希望表現の史的変遷―願望を中心に―」青木ほか編『コーパスによる日本語史研究 中古・中世編』ひつじ書房

- アノテーション付与の補助の可能性？
  - →古宮氏講演、近藤会長講演

研究対象資料の制約

- CHJのインフラ化により生じる課題
  - CHJ更新リスト：https://clrd.ninjal.ac.jp/chj/update.html
    - 「コーパスから漏れた抄物資料は、明らかに日本語史研究の表舞台から一歩下がったところに位置するものとなっている」（青木2022）
    - テキストデータベースも…
      - 岩波大系DB、噺本大系DBの休止（2023/4）
- 青木博史（2022）「抄物資料による日本語史研究の展望―歴史語用論の観点から―」『国語国文』91(11)
- 気付けば色々な変化が……
  - 今昔文節索引、表現社版虎明本…

研究対象資料の制約

- 「コーパスならではの研究」とは？
  - 国語研におんぶにだっこにならないようにするには？
- 「コーパスならではの研究」でなければ無意味か？
  - そういったデータを個人の研究者がデータ化する
  - 「表舞台」ではない資料についてのデータ化や情報の整備が、研究の多様性を保つための一つの方策になる
- 相対的に「一歩下がる」資料の利用可能性を高めていきたい

  →青池氏講演

「日本語資料に対する国立国会図書館のOCR 関連事業と成果物の活用」へのコメント

北﨑勇帆（きたざき ゆうほ）

高知大学人文社会科学部

NDLのOCR関連事業

- NDL Ngram Viewer
- 次世代デジタルライブラリー
  - OCRテキストデータ
  - OCRプログラム
    - NDLOCR
    - NDL古典籍OCR

次世代デジタルライブラリー

- 特に近代語・語彙史研究においては、
  「ある」ことを示すための必須のツールになるはず
  - （cf. 口頭発表B-5）
  - 「ない」ことは示せなくても、蓋然性は高くなる
  - 数を示すのには少し留意点が必要（→次項）
- 実は明治～昭和の叢書も多くヒットする
  - 日本名著全集、有朋堂文庫、人情本刊行会、漢籍国字解全書、…
  - 底本・校訂の問題はあるものの、やはり使わない手はない

次世代デジタルライブラリー

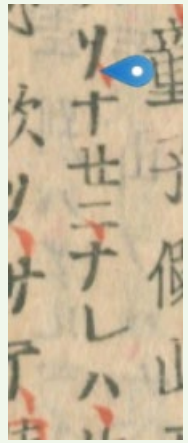
- 資料の発見にも使える
- 近代の高知方言小説を探してみる
  - 室津鯨太郎『南国』南人社、1926 https://dl.ndl.go.jp/pid/1019253
  - 『そりやえいものがありましたのうし、何時鯨が捕れましたぞのうし、御註進の来たのも知りませざつたよ』
  - 『今朝三津で捕れたとよ、早う知つちよつたら、己［おら］も買うがぢやあつたけんど』

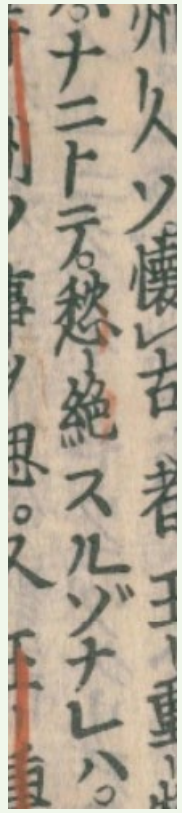
次世代デジタルライブラリー

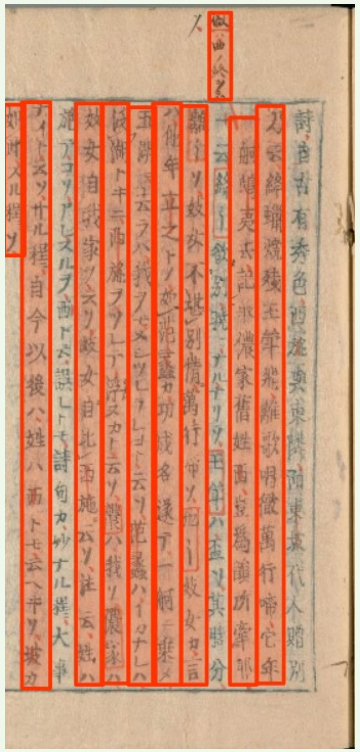
- 子供の作文の資料性
  - 「早く寝なんと朝起きれないぞ」
    （「子猫の行衛」学習研究会編『伸びて行く』6(1)、目黒書店、1926）
    https://dl.ndl.go.jp/pid/1803410/1/120
  - 妹はいつの間にか起きて、「起きれ／＼。」と私のふとんをはくつた。
    （百田宗治編『全日本子供の文章』厚生閣、1937）
    https://dl.ndl.go.jp/pid/1221364/1/77

OCRデータの扱い方（抄物を事例に）

- 間違え方の癖を知っておく
  - 十廿ニナレハ
    →ナセニナレハ（四河入海73-92）
  - 十二トテ絶スルナレハ
    ナニトテ愁絶スルゾナレハ（三体詩素隠抄8-22）
  - レイアウトの失敗（四河入海87-41）
    - 別紙の挟み込みなどもあり
- 「等しく誤る」わけではないので、
  探す対象がどう解析され得るかを把握する必要有







NDLOCR・NDL古典籍OCR

- 手元でOCRをかける
  - NDLOCR：https://github.com/ndl-lab/ndlocr_cli
  - NDL古典籍OCR：https://github.com/ndl-lab/ndlkotenocr_cli
- Google Colab上で動作するアプリ（東京大学・中村覚氏）
  - NDLOCR：https://zenn.dev/nakamura196/articles/b6712981af3384
  - NDL古典籍OCR：https://zenn.dev/nakamura196/articles/59fe1c9e76de65
  - ポチポチやってるだけでできる！

まとめ

- 「ざっくり検索」が許される範囲と許されない範囲で、
- 許される範囲では、データの癖を知ること
- 許されない範囲では、耐え得るデータを効率的に作ること
- 求コメント
  - やってみた事例
  - やってみたい事例