NINJAL
National Institute for Japanese Language and Linguistics

大規模研究資源の構築・整備の評価
国語研のコーパスを例に

2019年7月28日

小木曽智信

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

国立国語研究所のコーパス

コーパスとは：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大規模なデータベース

- 『日本語話し言葉コーパス』(CSJ) 2004年
- 『現代日本語書き言葉均衡コーパス』(BCCWJ) 2011年
- 『日本語歴史コーパス』(CHJ)　2013年～
- 『国語研日本語ウェブコーパス』　2016年
- 『多言語母語の日本語学習者横断コーパス』2016年～

※現在は日常会話、方言などのコーパスも構築中

コーパス
開発センター
Center for Corpus Development, NINJAL

コーパス開発センター

NINJAL
National Institute for Japanese Language and Linguistics

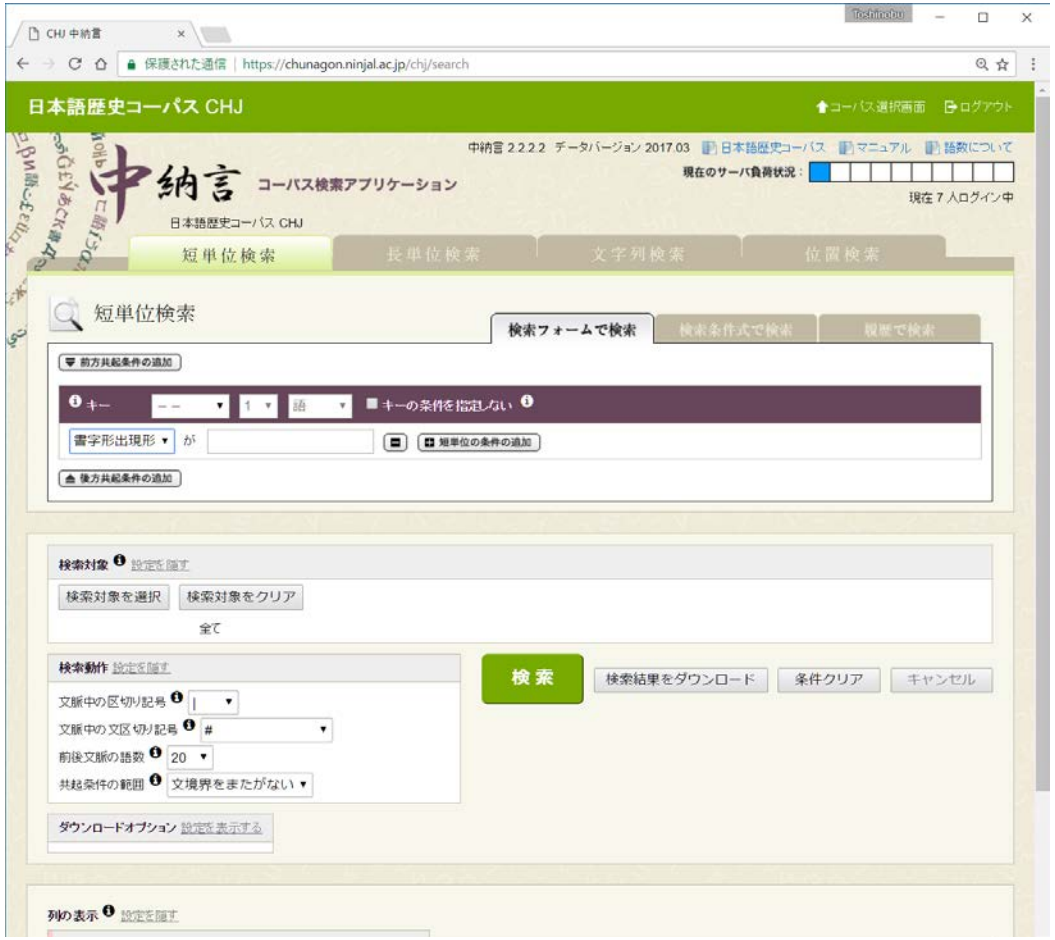
https://pj.ninjal.ac.jp/corpus_center/

研究インフラとしてのコーパス

- 「現代日本語書き言葉均衡コーパス」
  - 登録ユーザ数：約20,000人
  - 年間クエリ数：約40万件/年
  - 利用した論文数：約70本/年
- 「日本語歴史コーパス」
  - 登録ユーザ数：約10,000人
  - 年間クエリ数：約26万件/年
  - 利用した論文数：約50本/年（※予稿集含む）

コーパス検索アプリケーション「中納言」

- 人文系の研究者に利用しやすい形で提供
  - オンライン
  - 無料（要登録）



NINJAL
National Institute for Japanese Language and Linguistics

コーパス構築のコスト

「現代日本語書き言葉均衡コーパス」(BCCWJ)
2006～2010年(研究代表者:前川喜久雄)

- 約1億語の現代語書き言葉のコーパス
  - 書籍等からのバランスをとったサンプリング
  - 紙の文献の電子化
  - 単語情報の付与
- 予算:
  - 科研費 特定領域研究(総額 約8億円)
    ＋国語研運営費交付金

コーパス構築のコスト2

「日本語歴史コーパス」(CHJ)
2013年～現在(研究代表者:小木曽)

- 奈良時代から明治・大正時代までの日本語
  - 全本文に単語情報の付与
  - 外部の画像等にリンク

  ※2013年以前に構築されたデータを継承して含む
- 予算:
  - 国語研運営費交付金より 年間3000万弱
    ＋科研費 基盤A 年間1000万弱

コーパス公開のコスト

- サーバー代等：概ね年間1000万円程度
  （＋人件費、電気代 etc.）

- 新しいコーパスを構築する際は、その新規性をもって予算獲得ができるが、インフラとして定着したコーパスの公開・維持費は外部資金では困難

構築に携わる研究者のエフォート

- 構築担当の研究者はピーク時にはエフォートの大半（70%以上）をコーパス構築にあてる
- 論文にしづらい知的作業の結晶

➢ 研究・開発に費やしたエフォートに対する正当な評価が求められる

コーパスを研究業績とするときの問題点

1. 論文でないものを業績とすること
   - 引用等のありかた
   - 業績としての見せ方

2. 大規模データと個人の業績
   - 組織名しか表に出ない
   - 個人の業績としての見せ方

1. 論文でないものを業績とすること

引用のありかた

- コーパスは明らかに言語研究に不可欠な、広く利用されているものだが、論文等と違ってそのことの明示がされにくい

- 文献としての引用
- 言語資源（コーパス）そのものの引用

文献としての引用

- 従来の慣習に則り、参考文献としてコーパス開発時の研究論文を引用させる
  - 例：
    Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. "Balanced corpus of contemporary written Japanese". Language Resources and Evaluation 48 (2), pp.345-371 (DOI 10.1007/s10579-013-9261-0),2014:06.
  - コーパスのアップデートなどが反映されない
  - 必ずしも適切な論文があるとは限らない

言語資源そのものの引用

- 言語資源そのものを文献と同様の形で引用させる
  - 例:
    国立国語研究所(2011)『現代日本語書き言葉均衡コーパス』ver.1.0
  - 研究者の個人名が消えてしまう(後述)
  - Web公開の場合に永続的でなく、不安定

コーパスの利用規約では・・・

- 多くの場合、上記のいずれかを求めているが、あまり守られていない
  - 紙の文献であれば1冊ずつ資料を掲載するのに、利用した1億語のコーパスを挙げない

➢ 研究者、学界の意識改革とガイドラインの整備が求められる

2. 大規模データと個人の業績

個人業績の集積としてのコーパス

- どれだけ大規模であっても、コーパスは研究者「個人」の知的営為の成果の集積である
- 規模が大きくなると組織名だけが前面に出て「個人」が消えてしまう

- 構築に膨大なエフォートを費やすコーパスを、研究者個人が自身の業績として挙げられるようにすべき

「日本語歴史コーパス」での工夫

- コーパス構築に関わる多くが任期付き／非常勤。任期中に構築したコーパスを研究業績としたい

- 時代別に「サブコーパス」を設定し、構築参加者がいずれかの主担当となる
- 任期中に主担当サブコーパスを（いくつか）構築・公開することを任務とし、そのサブコーパスを研究業績とする

「日本語歴史コーパス」のサブコーパス

奈良時代	☑万葉集　☐宣命	
平安時代	☑仮名文学	☑和歌
鎌倉時代	☑説話・随筆　☑日記・紀行　☐軍記	
室町時代	☑狂言　☑キリシタン資料	
江戸時代	☑洒落本　☑人情本　☐近松	
明治・大正	☑雑誌 ☑教科書 ☐文学作品 ☐新聞 ☑明治初期口語資料	

CHJの出典情報（全体）

『日本語歴史コーパス』を利用した研究成果等を発表される際は、必ず下記の情報を明記してください。

- 国立国語研究所（2019）『日本語歴史コーパス』（バージョン2019.3，中納言バージョン2.4.2）https://chunagon.ninjal.ac.jp/（2019年4月1日確認）

- 国立国語研究所（2019）『日本語歴史コーパス』バージョン2019.3 https://chunagon.ninjal.ac.jp/

CHJの出典情報（サブコーパス）

- サブコーパス別に個人を明記した出典情報

国立国語研究所（鴻野知暁ほか）編（2017）『日本語歴史コーパス　奈良時代編Ⅰ万葉集』（短単位データ1.0 / 長単位データ 1.0，中納言バージョン 2.3）
https://pj.ninjal.ac.jp/corpus_center/chj/nara.html（2017年9月29日確認）

国立国語研究所（冨士池優美・須永哲矢・池上尚ほか）編（2016）『日本語歴史コーパス 平安時代編』（短単位データ 1.1 / 長単位データ 1.1，中納言バージョン 2.2.0）
https://pj.ninjal.ac.jp/corpus_center/chj/heian.html（2016年3月31日確認）

国立国語研究所（市村太郎・渡辺由貴ほか）編（2016）『日本語歴史コーパス 室町時代編Ⅰ狂言』（短単位データ 1.1 / 長単位データ 1.1，中納言バージョン 2.2.1）
https://pj.ninjal.ac.jp/corpus_center/chj/kamakura.html（2016年10月26日確認）

そのまま個人の業績としても掲載できる

まとめ

- 大規模研究資源の構築はコストもエフォートもかかる（が、そのぶん研究に活用されている）
  - 研究資源構築がきちんと業績として評価されるべき
- 大規模研究資源そのものは論文でないため適切に引用・評価されないことが多い
  - 意識改革やガイドラインの整備
- 大規模研究資源では研究者「個人」が消えてしまいがち
  - 個人業績にできるような工夫