NINJAL
National Institute for Japanese Language and Linguistics

現代語の書き言葉コーパスが果たす役割
―『現代日本語書き言葉均衡コーパス』の意義と今後の課題―

2023/07/21

文化審議会国語分科会国語課題小委員会

小木曽智信

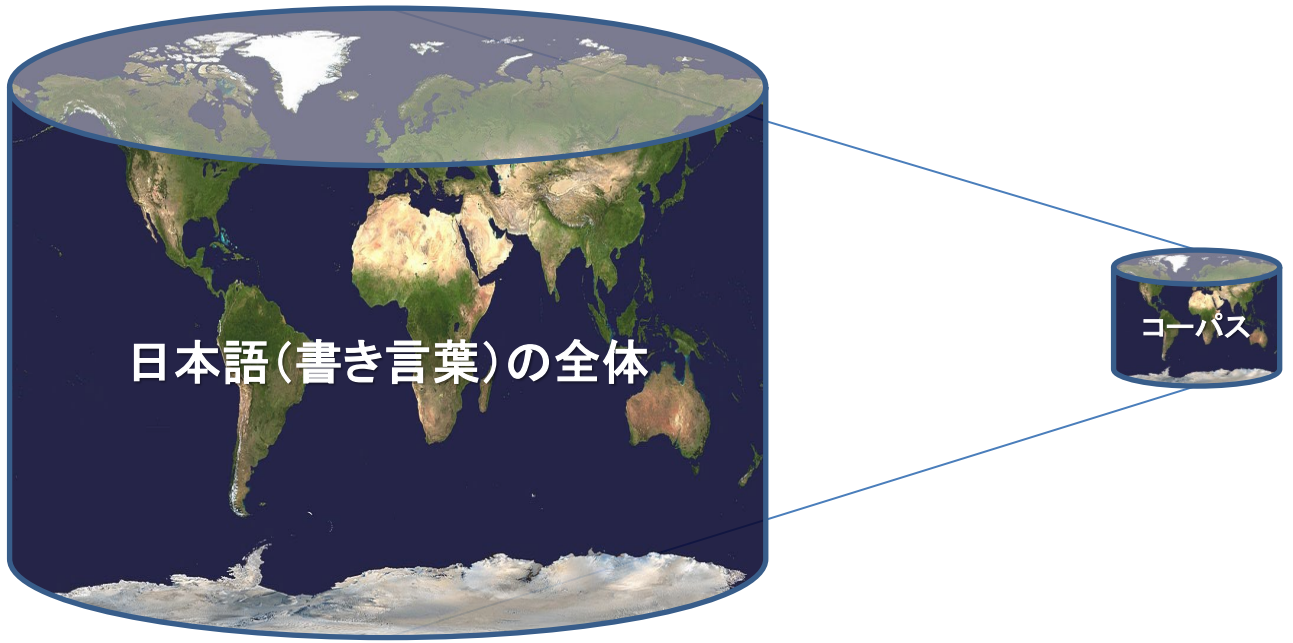
大学共同利用機関法人 人間文化研究機構
国立国語研究所
National Institute for Japanese Language and Linguistics

現代日本語書き言葉均衡コーパス

- 国立国語研究所が中心となって構築した初めての大規模なコーパス（約1億語）
- 2006年に構築を開始、2011年に完成
- 約1億語の現代語書き言葉を収録
- 全ての語に長短2種類の単語情報を付与
- 現代の書き言葉の代表性を有する「均衡コーパス」
- 略称、BCCWJ（Balanced Corpus of Contemporary Written Japanese）

均衡コーパスとは

- これを調べれば対象とする言語（現代日本語の書き言葉）の全体像が分かる（=代表性 representativenessを持つ）ように多様な言語資料をバランスよく収集したコーパス



BCCWJの3つのサブコーパス

生産実態を反映

出版サブコーパス

書籍、雑誌、新聞
2001～2005年
約3500万語

5年間に出版されたものが母集団

図書館サブコーパス

書籍
1986～2005年
約3000万語

図書館で広く流通した本が母集団

流通実態を反映

特定目的サブコーパス

白書、教科書、広報紙、ベストセラー
Yahoo!掲示板、Yahoo!ブログ、韻文、法律、国会会議録
対象期間はさまざま
3500万語

（母集団設定なし）

BCCWJの構築と予算

- リーダー：前川喜久雄（現・国立国語研究所長）
- 構築期間：2006～2010年
- 総額約16億円
  - –国立国語研究所の運営費交付金 ＋ 科研費

BCCWJ構築開始時に想定していた用途

表１．書き言葉均衡コーパスに想定される用途

日本語学	主観を排した言語分析 現代日本語の実態に即した文法・語彙の分析
日本語教育	基本語彙、基本構文、共起関係
国語教育	教育用基本語彙の選定
辞書編纂	用例収集、共起関係
心理学・認知科学	実験における言語刺激の統制
自然言語処理	統計的学習データ、アルゴリズム評価用データ
音声合成・認識	言語モデルの学習
国語政策	常用漢字の見直し、正書法の提案
文化資源	未来の文化財としての価値

前川喜久雄（2007）「特定領域研究『日本語コーパス―目標，進捗状況，そして夢―」特定領域研究「日本語コーパス」平成18年度公開ワークショップ（研究成果発表会）予稿集，pp.1-12 より

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
  - –ウェブ上で一般公開
  - –無料（要登録）

  - –単語情報による検索
  - –組み合わせによる高度な検索
  - –検索結果のダウンロード

https://chunagon.ninjal.ac.jp

NINJAL
National Institute for Japanese Language and Linguistics

研究・教育におけるBCCWJの利用状況

- 登録ユーザ数：約45,000人 ※ 2023年7月現在
- 年間クエリ数：約99万回/年 ※ 2023年7月までの1年間
- 利用した論文数：約180本/年 ※ 2022年刊行の既確認分
- 授業用アカウント発給数：126授業（全国の大学等、約4,200人分）
  ※2022年度

➢日本語研究のインフラとして重要な位置を占めている
（上記の数字は年々増加中）

産業界におけるBCCWJの利用状況

- 有償版BCCWJの商業利用契約数：約70件
  （一般契約は約50件，アカデミック版は約400件）
- 商業利用は1件あたり40～800万円
  https://clrd.ninjal.ac.jp/bccwj/fee.html
- 国際的な大手IT企業（いわゆるGAFAMを含む）数社とも利用契約を結んでいる

- 利用目的：言語解析器・検索システム・機械翻訳のモデル etc.

言語文化の記録としてのコーパスの役割

- BCCWJの構築時の目的の一つに「文化資源：未来の文化財としての価値」を挙げていた
- コーパスはその時の言語の実態を（いつでも確認できる形で）記録・保存する文化財ともいえる
  （BCCWJは2005年当時の日本語の記録）

- 意識調査とならぶ実態調査のための基礎資料となる
  –しばしば言語に関する意識調査と実態調査は結果が乖離する

コーパスが果たす役割

コーパスの主要な役割

1. 学術的価値：言語研究に欠かせない基礎データとしての役割
2. 産業的価値：情報処理（言語処理・AI）の学習用データとしての役割
3. 文化的価値：言語文化を継承・保存し発展させる基盤としての役割

BCCWJの限界と拡張の必要性

BCCWJの限界

1. 収録資料が現代語としては古くなっている
   - 主要な収録データは2005年まで。「スマホ」が出てこない
   - 現在の学生にとって「現代語」と言えなくなりつつある
2. 経年的な調査ができない
   - 2005年当時の日本語を切り取ったコーパスであり、現在進行中の変化を追いかけることができない
3. 規模が必ずしも十分でない
   - 研究対象によってはサイズが足りず、特に言語処理ではより大きいコーパスが望まれる
   - 他の言語のコーパスはより大規模なものが多い

他言語の主な現代語コーパス

- イギリス：British National Corpus 約1億語
- アメリカ：Corpus of Contemporary American English 約10億語
- ドイツ：Deutsches Referenzkorpus (DeReKo): 約550億語
- スペイン：Corpus del Español：約20億語
- ロシア：National Corpus of Russian Language: 約4億5千万語
- ポーランド：National Corpus of Polish (Narodowy Korpus Języka Polskiego): 約10億語

多くのコーパスが定期的に更新・データ追加されている

※語数は概数（時期により異なる）

他言語の主な現代語コーパス(2)

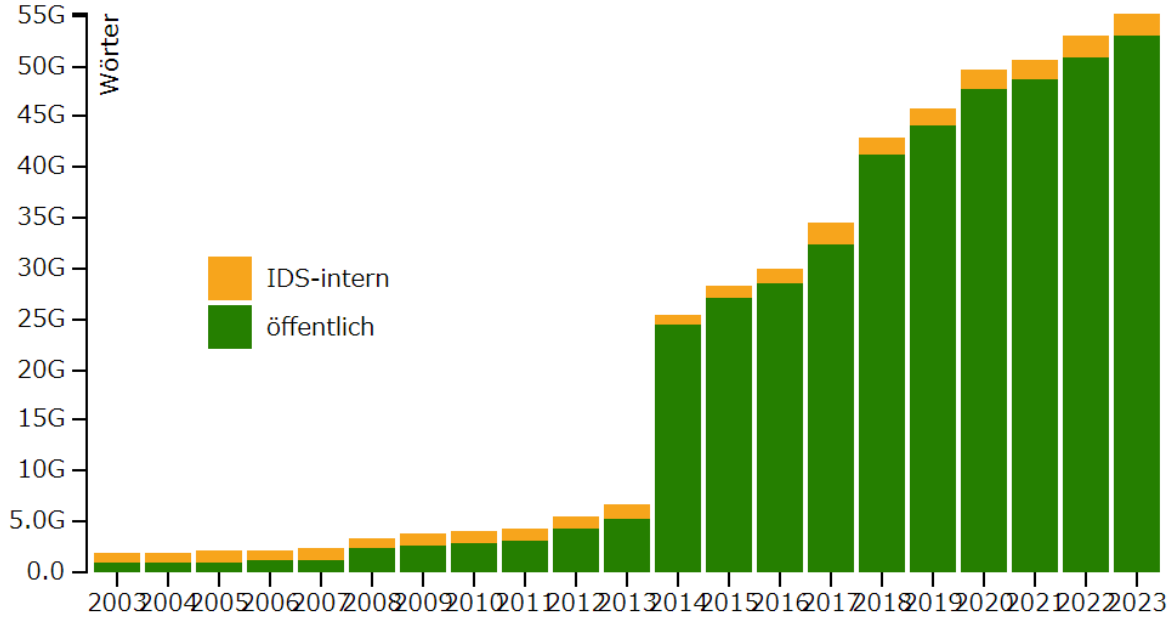
言語文化の保存・振興の観点から比較的話者の少ない言語のコーパス構築も盛ん

- チェコ：Czech National Corpus (Český národní korpus): 約10億語
- スロベニア：GIGAFIDA: 約10億語
- ウェールズ：National Corpus of Contemporary Welsh (Corpws Cenedlaethol Cymraeg Cyfoes)　：1千万語

※海外のコーパスの一部については専修大学・丸山岳彦教授から情報提供をいただきました

Deutsches Referenzkorpus (DeReKo)の例

- ドイツ語の現代語コーパス
- 定期的に更新されどんどん大きくなっている
- 経年変化も追える
- ただし均衡コーパスではない



https://www.ids-mannheim.de/digspra/kl/projekte/korpora/archiv-1/

BCCWJの拡張・発展の必要性

1. 収録資料が現代語としては古くなっている
   →最新のデータを収録して更新したい
2. 経年的な調査ができない
   →経年調査ができるよう継続的に拡張していきたい
3. 規模が必ずしも十分でない
   →BCCWJに加える形で全体として規模を大きくしたい

これができれば極めて高い文化的価値と応用可能性を持つコーパスになる

コーパス構築のコストの問題

- 構築期間：2006～2010年
- 総額約16億円
  –国立国語研究所の運営費交付金 ＋ 科研費

- コーパスの拡張・発展だけでは学術的新規性は乏しいため、今後も研究費として予算を獲得することは難しい
- 他国でも見られるようにコーパスの文化的価値にも着目した、行政的なサポートによる継続的拡張ができることが望ましい

大規模言語モデルとコーパス

大規模言語モデルとコーパス

- ChatGPTなどで知られる大規模言語モデル（LLM）の構築にもコーパスが使われるが、ここで用いられるコーパスのサイズは文字通り桁違いに（4～6桁くらい）大きい
- BCCWJの約1億語（テキストは0.5GB程度）に対し、OpenAIのGPT-3は570GB（同様のテキストなら1000億語以上）、GPT-4は非公開だがさらにずっと多いと言われる

- ただし、これらLLMで用いられるコーパスはウェブから集められたデータで日本語はせいぜい5%程度しか含まれない

大規模言語モデルとコーパス(2)

- 最近になって国内で日本語に特化したLLMが開発、公開されている（rinna japanese-gpt，CyberAgent OpenCALM, LINE HyperCLOVA, NICT など）
- これらの日本語の学習コーパスもウェブから集められた大量の日本語データ
  - –Rinna, OpenCALMは Wikipediaやcc100等
  - –LINE（LMコーパス）は11.8TB（5000億語）
  - –NICTは350GBのウェブ日本語テキスト

大規模言語モデルとコーパス(3)

- ウェブから得られるコーパスはサイズは大きくても内容に偏りがあり、場合によっては著作権や個人情報の問題もある
- 今後はAIが生成したウェブ上のテキストをAIが学習に使ってしまうデータ汚染、学習用のデータに悪意ある改ざんを行って攻撃するデータポイズニングの危険も指摘される

- ウェブから離れた所から(も)集められたBCCWJのようなコーパスは大規模言語モデル時代にも役に立つはずである

大規模言語モデルとコーパス(4)

- BCCWJのように高品質で安心して利用できるコーパスは、比較的小さいサイズであっても今後の大規模言語モデルに貢献できる可能性がある
  - –ファインチューニング（再学習による調整）や評価用のデータなど
- 教科書的な高品質のテキストを用いることで小さなモデルでも高い性能を引き出せるという新しい研究もある
  - –Suriya Gunasekar et.al (2023) Textbooks Are All You Need https://doi.org/10.48550/arXiv.2306.11644

おわりに

- 『現代日本語書き言葉均衡コーパス』を例に、コーパスが果たす役割を研究・産業・文化の面で見てきた
- 今後もこうしたコーパスの役割を発展させていくためには、コーパスを継続的に拡張していくことが望まれる
- 意識調査と実態調査は言語調査の両輪であり、今の日本語をコーパスとして継続的に記録・保存して実態調査を行うことを文化行政の中で進められるなら、国立国語研究所はこれまでのノウハウを活かして協力することができる