

2024/08/01 文化庁文化審議会国語分科会言語資源小委員会第2回会議報告資料

継続・普及・実装：日本語言語資源の未来に求められる3つの方向

—英語コーパス研究からの提言—

石川慎一郎（神戸大学教授）

・本報告は、報告者個人の見解に基づくまとめであり、所属する組織・機関・学協会の見解を反映したものではありません
・著作権配慮のため、公開資料では、図版等を削除します
・本資料の公開後に見つかった誤植・誤記については、報告者のウェブサイト上で修正を示します

1. 日本語言語資源に必要な3つの観点

- ・継続（データを取り続けていく）
- ・普及（集めたデータが広く普及するよう、検索システムなどの工夫を行う）
- ・実装（データの社会実装の道筋を示す）

2. 継続的収集：言語変化を記録するために

- ・コーパスは、言語母集団に対する「標本」。標本抽出のためには母集団の同定が必要で、通例、特定の年（代）の言語を母集団とみなす。ゆえに、標本コーパスは、言語変化のある瞬間を切り取った「スナップショットコーパス」となり、完成した瞬間にすでに古くなっている。■ 言語変化を追跡していくためには、同一基準に基づくデータの継続的収集が不可欠
- ・COCA（1990-2019の30年間、同一基準でウェブ上の米語資料を継続収集）のように、データを継続的に収集したコーパスがあれば、個々の語や表現の使用度や用法の変化が科学的に検証できる
- ・データ継続収集のいくつかのパターン
 - （1）同一機関によるデータ継続収集
- ・British National Corpus 1994 ■ British National Corpus 2014
- ・いずれのコーパスも英国ランカスター大学が関与

- ・実際の責任者は故 Geoffrey Leech 氏から Tony McEnery 氏に代わっている
- ・コーパス開発は属人的な部分が多く、長期的予算がないと、大学・研究所といえども、人材の転出・退職などで、コーパス構築のノウハウが失われることも

(2) 同一個人による継続的収集の例

- ・Corpus of Contemporary American English (COCA)
- ・元 Brigham Young University 教授の Mark Davies 氏が個人で構築
- ・2008 年から構築を開始。その後、2020 年に氏は BYU を退職。
- ・当初、所属機関名を冠して BYU Corpus と呼ばれていたが、退職後、English-corpora と改称される
- ・今回はたまたま継続がうまくいったが、大学→個人のデータの引継ぎには様々な問題も予想され、所有権と管理責任の所在があいまいになって、データの収集やメンテナンスが滞る場合も

(3) 拡張分散型による継続的収集の例

- ・Brown Corpus (米国 Brown University で、1961 年の米国の書き言葉を収集・公開)
- ・独自のサンプリングフレームに基づき、2000 語のサンプルをジャンル別に 500 本集めて 100 万語にする
- ・その後、様々な第三者が、同じ枠組を用い、類似のコーパスを構築
- ・アメリカ英語: B-Brown, Brown, Frown, AmE06, Crown ほか
- ・イギリス英語: BLOB, LOB, FLOB, BE06, CLOB ほか
- ・その他の英語: Kolhapur (インド), ACE (豪), Wellington (NZ) ほか
- これらを総称して “Brown Corpus Family” と呼ぶ
- ・「手ごろなサイズと、明瞭で踏襲しやすい枠組み」があれば、第三者による自発的なコーパス拡張が可能になる場合も

本節のまとめ

- ・コーパスづくりは、実質的には、機関でなく個人が担う場合が多い
- ・当該教員が転出・退職すれば、機関としてのコーパス開発はそこで終わる
- ・個人としての継続も大学退職後は難しい
- ・数年単位の競争的資金ではなく、10年・20 年のスパンでの長期支援があってはじめて、大学や研究機関は、データを集めると同時にそのノウハウを受け継ぐ人材を養成していける

3: 普及: 使いやすい検索インタフェースの開発

- ・検索インタフェースの基本機能= KWIC (key word in context): 検索語を含む行を同時に表示。並べ替え (sorting) を行うことで、当該語の前後の共起パターンなどを直観的に把握できる。
- ・KWIC は、用例からパターンを抽出するトレーニングを積んでいる職業的言語学者・辞書編纂者以外には、使いこなすのが難しい

・海外のオンラインやスタンドアロンのコンコーダンスには、コーパスデータからわかることを専門家以外にもわかりやすく伝えるための工夫が凝らされている

(1) English-corpora.org の検索インタフェースの場合

- ・共起語の一括表示 (棒グラフで頻度を直観的に示す)
- ・ジャンル別・時代別頻度の棒グラフでの一括表示
- ・類義語の共起語の差を一括表示 (統計的に有意な差があるものをハイライト)
- ・関連情報の一括集約 (ジャンル別頻度、成句、類義語、ジャンル別頻度、用例など)

(2) LancsBox の場合

- ・単語コロケーションネットワーク表示

(3) BNClab の場合

- ・性・年齢・地域・階層・時代・産出モード別に頻度差を一括表示

本節のまとめ

- ・言語資源に対する言語学者のニーズと、教育関係者 (教師、学習者)、また、広く一般国民のニーズは同じとは限らない
- ・言語学者が重視する KWIC 出力は、大量の用例を咀嚼・分類・解釈し、そこから抽象的なパターンや傾向を抽出する高度な言語解析スキルを前提とする
- ・コーパスをさらに普及させていくためには、言語学分野外の関係者とも意見を交わし、社会が求めるデータを、社会が求める形で提供する方法 (新しい検索インタフェースの開発など) について考えていく必要もあるだろう

4: 社会実装: 教育や社会の問題解決に役立てる

(1) ESL 辞書への応用

- ・John Sinclair (1987) Collins COBUILD English Dictionary
- ・コーパスから得られた「本物の英語」を英語学習者に提供
- ・BNC1994 のリリースを受け、LDOCE、OALD 等の主要辞書もすべてコーパス準拠に転換

(2) 社会における言語規範の更新

- ・BNC2014 Spoken のリリースによる The Telegraph の 2017/9/24 記事
- ・Dr. Claire Dembry (CUP) 「英語教育はコーパスでしか観察できない言語変化を反映すべき」

(3) 言語教材の具体的な改良

- ・McCarthy, McCarten, & Sandiford (2006), *Touchstone* (Cambridge UP)
- ・Cambridge International Corpus の解析に基づく英会話教材

(4) 社会問題解決への応用

- ・Brookes & Collins (2024) *Corpus Linguistics for Health Communication: A Guide for Research*

・”obesity”の共起語を調査→obesity についての社会的認識やステレオタイプを解明し、治療に役立てる

・医師／看護師と患者の会話と一般の会話を相互比較し、医療従事者の言説を分析

本節のまとめ

・英語圏において、コーパスの社会実装は徐々に進んでいる

・コーパスに準拠した辞書・教材・テストの開発

・母語の文法規範の見直しや、「正しい英語」についての社会的論争のきっかけの提供

・（本日は触れる時間がなかったが）学習者コーパス (learner corpus) を用いた外国語としての言語教育の新しい枠組みの提案なども

・言語以外の分野、たとえば医療や公衆衛生分野においても、コーパス活用が始まっている

・新聞コーパス分析による疾病をめぐる社会的ステレオタイプの解明

・医療者発話コーパスの分析に基づく、望ましい医療者の話し方の解明、患者の心理変化をもたらす有効なアプローチの定式化、など

5.言語資源研究に必要なサポートの例

(1)コーパス開発を手掛ける主要な研究機関への継続的支援

・コーパス開発は、本来、属人的。これを「システム」にするには、人の入れ替わりも見据えた 20 年スパンでのプロジェクト支援が必要

(2) 評価されにくい言語資源の整備を促進するための施策

・「公開型言語資源／関連ツール開発・維持」を科研分野に新設するなど

(3) 共有型の検索インタフェースの構築

・利用者が自由にデータをアドオンできる高性能インタフェースの開発・公開（参考：英国の CQP web など）

(4) データ入手の支援

・医療・保健・工学・政治分野など、個人研究者では集めにくいデータを体系的に記録・利活用するための仕組みの整備