

2024/11/25 文化庁文化審議会国語分科会言語資源小委員会第4回会議報告資料

言語資源の社会実装 ―「集めた後」を見据えて―

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、図版等を削除しています。
- ・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. (前回報告)言語資源開発に必要な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. 言語資源の社会実装を考える必要性

- ・「コーパスに価値がある」ことは言語学者には自明
- ・しかし、なぜそれが重要なのか、どんな利点があるのかを広く国民に知らせていく努力も必要

3. 言語研究を社会に開く―応用言語学の方向性―

- ・理論か応用か →「実際の・日常的課題解決を志向する」応用言語学
- ・言語学の応用と応用言語学
- ・言語→社会へ: 理論系言語学、用例重視の実証的言語学、社会志向型言語研究、社会実装型言語関連研究、社会思想系言語研究の位相

4. CASS の取り組みに学ぶ

・英国の UKRI の一端をなす ESRC の支援による ESRC Centre for Corpus Approaches to Social Science (CASS)

・4大ミッション

- 1) コーパスアプローチの利点を幅広い社会科学分野に提供する
- 2) コーパス言語学に専念するのではなく、言語分析の最新技術を社会科学のさまざまな問題に応用する
- 3) ESRC による 5 年間の投資を基盤として、今後 15 年間にわたって、コーパス技術に長じた次世代社会科学研究者を育成
- 4) 社会科学におけるコーパス技術普及を促進する

・主な研究方向： 1) 公衆衛生、2) 経済ビジネス、3) 社会心理社会思想、4) 国内国際政治、5) 文学・言語・言語教育

・CASS の取り組みから学べること

- 1) 大規模汎用コーパス (BNC、BCCWJ) だけで社会実装を行うのは難しい
- 2) コーパスやコーパス言語学そのものではなく“corpus approach”を接点とするプロジェクト方式を通して、他分野の研究者を意図的に巻き込んでいく
- 3) 各種の社会的課題(差別偏見解消、教育改革、行政システム効率化など)の解決を目指す様々な小規模コーパスの開発・分析を支援
- 4) そうして作られる多様な社会課題特化型コーパスネットワークの「絶対参照基準」として大規模汎用コーパスの整備・維持を図る

5. さまざまな実践例

- ・ASD 診断支援
- ・抑鬱症診断支援
- ・肥満症治療支援
- ・医療専門職者向けの語学トレーニング支援
- ・学習者に特化した語彙選定
- ・L2 能力の簡易診断法の開発

6. ICNALE プロジェクトから考える《継続・普及・実装》

- ・アジア圏国際英語学習者コーパス ICNALE (2007-現在まで進行中) ■継続的にメンテナンスと拡張
- ・普及のために、専門家向けダウンロード版公開と、一般向けオンライン検索システム開発・公開の2トラックを用意
- ・オンラインシステムの概要 (KWIC、共起語、特徴語、音声、動画、校閲、頻度グラフなど)
- ・研究者への普及 (英語学・英語教育学・自然言語処理・自動認識・自動誤用訂正・AI研究など)
- ・社会実装の道筋 (文法チェック、shared task、辞書開発)
- ・コーパスに入ることによって世界の研究の対象になる、という側面も

7. 小括 ―報告者による提言―

- (1) 言語資源は作るだけでなく、完成後もアップデートを行い、かつ、普及のための工夫を凝らすことで、社会実装まで道筋をつけることが重要
- (2) 応用言語学が掲げる「社会的課題解決のための言語研究」という理念への戦略的接近を図る
- (3) コーパスそのものよりも、むしろその分析技術を接点として、戦略的に他分野を巻き込んでいく (とくに、医療、社会福祉、経済、司法など)
- (4) 他分野のデータを使った各種の小規模コーパスの作成・解析を積極的に支援するとともに、それらの《絶対参照点》として大規模な基盤的日本コーパスの整備を位置づける
- (5) 長期にわたる安定した公的資金投入によって、基盤コーパス整備とともに、「社会実装志向型異分野共創言語資源活用研究」の普及を図る

参考文献

- Batool, H., Rasool, S., & Shehzad, W. (2022). Content words and their conceptualization: A corpus-based study of conversations of children with autism. *Hayatian Journal of Linguistics and Literature*, 6(1), 207-230.
- Collins, L., Baker, P., & Brookes, G. (2024). Representations of obesity in Australian and UK news coverage: A diachronic comparison. *Applied Corpus Linguistics*, 4(2), Article 100092.
<https://doi.org/10.1016/j.acorp.2024.100092>.

- 大学英語教育学会基本語改訂特別委員会 (2016) 『大学英語教育学会基本語リスト新 JACET8000』 桐原書店.
- 石川慎一郎 (2021). 「絵描写作文課題における L2 日本語学習者の動詞使用と習熟度の関係 —I-JAS の SWI 課題データの計量的概観—」 『統計数理研究所共同研究レポート』 444, 1-22.
- 石川慎一郎 (2023). 『ベーシック応用言語学第 2 版: L2 の習得・処理・学習・教授・評価』 ひつじ書房.
- Ishikawa, S. (2023). *The ICNALE guide: An introduction to a learner corpus study on Asian learners' L2 English*. Routledge.
- Kato, S., Hanawa, K., Saito, M., & Nakamura, K. (2024). Creating a diagnostic assessment model for autism spectrum disorder by differentiating lexicogrammatical choices through machine learning. *PLOS ONE*, 19(9), e0311209. <https://doi.org/10.1371/journal.pone.0311209>
- 岸野英治(編)(2019). 『ウィズダム和英辞典 3 版』三省堂.
- Staples, S. (2015). Examining the linguistic needs of internationally educated nurses: A corpus-based study of lexico-grammatical features in nurse-patient interactions. *English for Specific Purposes*, 37, 122-136.
- Staples, S. (2019). Using corpus-based discourse analysis for curriculum development: Creating and evaluating a pronunciation course for internationally educated nurses. *English for Specific Purposes*, 53, 13-29.
- 鈴木あすみ・幕内充・和田真・中村仁洋・石井亨視・小磯花絵(2024). 「日本語を母語とする自閉スペクトラム症者の公開コーパス構築」 『言語資源ワークショップ発表論文集』 1, 113-131.
- Tasnim, M., Ehghaghi, M., Diep, B., & Novikova, J. (2023). DEPAC: A corpus for depression and anxiety detection from speech. In Zirikly, A. et al. (Eds.) *Proceedings of the eighth workshop on computational linguistics and clinical psychology* (pp. 1-16). Association for Computational Linguistics.