国立国語研究所学術情報リポジトリ

はじめに

メタデータ	言語: Japanese 出版者: 研究社 公開日: 2026-07-24 キーワード: 作成者: 小磯, 花絵 メールアドレス: 所属: 国立国語研究所
URL	https://doi.org/10.15084/0002000718

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

はじめに

本書は、『日本語日常会話コーパス』(CEJC) の構築経験にもとづき、会話コーパスの設計・構築・利活用法をまとめたものである。

CEJCの特徴は、多様な場面における多様な話者による日常会話をバランスよく格納し、動画や複数のアノテーションを含めて公開する点にある。日常会話をバランスよく収録するためにコーパスをどのように設計すべきか、動画を含めて公開するにあたりいかなる法的・倫理的な問題が生じるのかなど、準備段階から数多くの問題に直面した。また日常会話を文字化し種々のアノテーションを施すには、書き言葉や講演音声とは異なる性質を持つ日常会話の特性を踏まえ、基準を慎重に設定する必要があった。

これらの経験から得られた知見を踏まえ、本書の前半 (第1〜4章) ではCEJCの設計・構築のプロセスを体系的にまとめた。コーパスを適切に活用するには、コーパスの設計思想や基準を理解することが不可欠である。またこうした記録は、新たに日常会話コーパスやマルチモーダル言語資源の構築を目指す研究者にとっても有用であろう。その点を考慮し、コーパスの設計や構築の中核的要素については、CEJCで採用した方法に限定せず、関連する情報を広く紹介するようにした。

第1章では会話コーパスの設計と構築の全体像を示す。話し言葉コーパスをいくつか取り上げ、設計面を中心に特徴を紹介した上で、CEJCを例に収録からアノテーション付与に至る構築過程を概説する。

第2章では会話の収録法と収録後の音声・動画データの処理を、第3章では音声の転記方法を扱う。これらは会話コーパス構築の中核であるため、CEJCの手法を参照しつつ、より一般的な観点から解説する。

第4章ではCEJCで新たに導入した談話行為情報の設計方針と体系について述べる。既存の枠組みを参考に、日本語の日常会話の特性に即して再構成したものである。

CEJCには談話行為情報のほかにも形態論情報や韻律情報などが付与されているが、これらは既存のコーパスにも付与されており解説も多いことから、本書では独立した章は設けない。ただし、形態論情報については第5章で、韻律情報については第7章で触れる。

後半の第5章から第8章では、CEJCを利用するための実践的方法を紹介する。第5章で取り上げる「中納言」は、短単位・長単位情報を用いた検索に適し、CEJC利用の第一歩として推奨される。登録は必要であるが無償で利用できる。第6章で紹介するELANは動画・音声を統合的に扱うことのできるマルチモーダルアノテーションツール、第7章で紹介するPraatは音声分析に適したツールである。これらは動画や音声を参照しながら精緻な分析をする際に有用である。第8章で取り上げるリレーショナルデータベース（CEJC-RDB）は、CEJCに付与された複数のアノテーションやメタ情報を統合的に扱う際に力を発揮する。動画・音声データやELAN・Praat用の専用ファイル、CEJC-RDBの利用には契約が必要であり、操作も「中納言」より複雑ではあるが、中納言を超える分析を志す読者にとっては有力な選択肢となるであろう。

本書が、会話研究のさらなる発展や新たな言語資源の創出に向けた一助となれば幸いである。

2025年11月

編者
小磯花絵