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 NINJAL
2022オープンハウス

『BTSJ日本語1000人自然会話コーパス』と『自然会話リソースバンク（NCRB）』

宇佐美まゆみ（国立国語研究所研究系 教授）

1.『BTSJ日本語1000人自然会話コーパス』とは

●宇佐美まゆみ監修（2022）**『BTSJ日本語1000人自然会話コーパス』**とは，シナリオのない自発的な自然会話を，母語場面，接触場面の初対面会話，友人同士の会話，教師と学生の論文指導場面等のサブグループごとに，年齢や性別を条件統制して収集した会話データをまとめたものである。

●『BTSJ日本語1000人自然会話コーパス』は，2002年の東外大時代から拡充を進め，2022年に国語研から公開した『BTSJ日本語自然会話コーパス（トランスクリプト・音声）2022年3月NCRB連動版』に**40会話を追加**し，全会話のうち**167会話については動画も追加**し，さらに，**NCRB（自然会話リソースバンク）と連携**をはかった，**「語用論的分析に適した」ユニークな自然会話コーパス**である。

2. 本コーパスの特徴と意義

●**シナリオのない，自発的な相互作用としての自然会話を集めた国内外最大級のコーパス**

①「言語社会心理学的アプローチ」（宇佐美1999），「総合的会話分析」（宇佐美2008）の方法論に基づき，会話参加者の年齢，性別，話題などを統制して収集し，「録音された会話以外の社会的要因」の分析も重視する。そのため，各会話グループのデータ収集条件や話題，話者の年齢・性別・職業，その他の属性の情報も提供するので，**計量的分析が可能**

②発話の重なりや沈黙など，多くのコーパスでは提供されていない語用論的分析に不可欠な情報を記してあるので，**定性的分析も可能**

③**「基本的な文字化の原則」**（BTSJ: Basic Transcription System for Japanese）によって文字化したトランスクリプトを収録することによって，各研究者が独自の観点から分析項目を設定し，コーディングできる。

④コーディングした分析項目の集計が特別のプログラミング能力がなくてもできるように開発したツールである**『BTSJ文字化入力支援・自動集計・複数ファイル自動集計システムセット』**を連動させている。システムセットは，**講習を受けると受領できる**。

●初対面と友人の会話データが豊富で，その話し方の違いが明確であるので，**相手によって言葉を使い分ける対話システムの開発にも応用**できる。また，母語場面（母語話者同士）と接触場面（母語話者と非母語話者）の会話の両方が格納されているので，**日本語教育にも応用**できる。

3. 本コーパスに収録されているデータ

●本コーパスに収録されているデータの情報一覧はExcel ファイルで，4つのシートで構成されている。会話データの概要よりも詳しい，データ収集者の会話収集の目的（例「いかに反論するかの男女差の検証等」）や，会話収集条件などの情報も記載されている。

①「会話グループ情報」シート…会話の収集方法や条件等の情報
②「個別会話情報」シート…会話の時間や話者数などの情報
③「話者情報」シート…話者の性別，年齢，出身，社会的属性などの情報
③「話者情報の注（話者記号の説明）」シート…話者記号の意味の説明

4.「完成版」コーパスの特徴

● 2022年3月に公開した「NCRB連動版」に対して，新たに**40会話を追加**した。「完成版」に収録されているデータの内訳は，下表のとおり。

●1会話20分程度の会話が，**514会話**（合計約127時間）収録されている。その主な基礎統計情報は，コア会話数296会話，非コア会話数218会話，**話者数1028人**（延べ人数）である。

●今回**初めてNCRB（自然会話リソースバンク）上で動画を公開**する。これをもって，**BTSJコーパスとNCRBの連携が完成**する。

●動画データを利用したい場合は，NCRB内の「動画データ申込フォームダウンロード」からフォームをダウンロードして必要事項を記入し，管理事務局の審査を受けることで，利用可能となる。

表1 BTSJコーパスに収録されているデータ（トランスクリプト・音声・動画）およびNCRBに収録されているデータ（動画）

データ	会話数	時間
トランスクリプト	130会話	39時間4分2秒
トランスクリプト+音声	217会話	43時間49分17秒
トランスクリプト+音声+動画	167会話	44時間16分7秒
合計	514会話	127時間10分26秒

『BTSJ日本語1000人自然会話コーパス』をNCRBに搭載

1. 自然会話リソースバンク（NCRB）とは

●**自然会話リソースバンク（NCRB: Natural Conversation Resource Bank）**とは，「自然会話データを使った研究」と「自然会話を素材とする教材」の二つの部分で構成される，共同構築型多機能データベースである。

●共同構築型のため，既存の説明に追加や修正が可能で，その場合，直前の説明記入者あてには変更の通知がメール送信される。

●**「研究部門」**には，自然会話の実態を分析するために必要なデータ（スクリプト・音声・動画）が搭載されており，それらを利用して自然なコミュニケーション能力を養成するための基礎研究を行うことができる。

●**「教材部門」**には，様々な日本語学習者の学習ニーズに応じたコンテンツ（「会話教材」）が搭載されており，それらを利用し「自然会話を素材とするWEB教材」を作成し，日本語教育の現場で利用することができる。

2. 研究部門の特徴

●これまでは「教材部門」を先行開発したが，今回**新たに「研究部門」も完成**した。

●『BTSJ日本語1000人自然会話コーパス』に収録されている**514会話（うち167会話は動画データ付き）**が搭載されている。

●「データを登録したい方はこちら」という入口からは，『BTSJ日本語1000人自然会話コーパス』のデータを視聴でき，自身のデータをアップロードしたり，他の研究者と共有したりすることもできる。自身の研究テーマに合わせて**複数の会話をグループ化**することも可能である。

●「データを利用したい方はこちら」という入口からは，『BTSJ日本語1000人自然会話コーパス』のデータを，自身の研究内容に合わせて**個別に会話をダウンロードしたり，グループ化した会話を一括してダウンロードしたり**できる。

3. 教材部門の特徴

●「自然会話を素材とするWEB教材」とは，シナリオのない「自然会話」をそのままWEB教材化したものである。共同構築型WEB教材は，**世界初の試み**である。

●詳細は，昨年のポスター（下記URL）をご覧ください。
https://www2.ninjal.ac.jp/openhouse/2021/pdf/b11.pdf

本研究は，国立国語研究所第3期機関拠点型基幹研究プロジェクト「日本語学習者のコミュニケーションの多角的解明」，サブ・プロジェクト「日本語学習者の日本語使用の解明」（リーダー：宇佐美まゆみ），科研基盤研究(A)「語用論的分析のための日本語1000人自然会話コーパスの構築とその多角的研究」（研究代表者：宇佐美まゆみ）の成果の一部である。