

## 『BTSJ 日本語 1000 人自然会話コーパス（完成版・映像付き）』の 特徴と活用法

宇佐美まゆみ（国立国語研究所）・小川都（国立国語研究所）

### 1. はじめに

本発表では、2021 年度に公開予定の『BTSJ 日本語 1000 人自然会話コーパス（完成版・映像付き）』（以降、BTSJ コーパス）を紹介し、その特徴と活用法について解説する。本コーパスは、これまで拡充・整備を続けてきた『BTSJ 日本語自然会話コーパス（トランスクリプト・音声）2021 年 3 月版』（446 会話、112.5 時間）（延べ話者数 892 人）にさらに 76 会話（延べ話者数 175 人）を追加し、全 522 会話、約 137.7 時間（述べ話者数 1067 人）を公開するものである。「**1000 人日本語自然会話コーパス**」の構築を目標とするプロジェクトとして企画されたもので、これまで非公開であったものも含めて、今回初めて「映像データ」も公開するものである。

### 2. 『BTSJ 日本語 1000 人自然会話コーパス（完成版・映像付き）』の特徴

本コーパスの独自の特徴は、大きく分けて、以下の 4 点である。①「会話フォルダ」ごとに条件が統制されて収集されたデータの『基本的な文字化の原則（BTSJ: Basic Transcription System for Japanese）2019 年改訂版』（宇佐美, 2019: 以降 BTSJ と呼ぶ）によるトランスクリプト、音声、映像が含まれていること、②『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』と連動していること、③共同構築型多機能データベース「自然会話リソースバンク NCRB (National Conversation Resource Bank)（以下 NCRB と呼ぶ）」に格納されており、映像データ付きのものは、NCRB 上の「自然会話を素材とする Web 教材」とも連動し、NCRB の「教材作成支援機能」を用いて、「動画、トランスクリプトに基づく解説、Q&A」がセットになった「Web 教材」が、必要事項を入力していくだけで作成できる点。④最後の、そして最も重要な点は、本コーパスが、③の NCRB という「共同構築型」多機能データベース NCRB というプラットフォームに格納されることによって、**21 世紀型「共同構築型」コーパス**という流れを企画し、世界で初めての試みとして実践していくという点である。以下、順に説明していく。

#### 2.1 「会話フォルダ」ごとにまとめられたデータの特徴について

「自然会話データ」は、書き言葉と比べてデータ収集に膨大な時間と労力がかかるため、個人、或いは、少人数の研究グループだけでは、多くの会話データを扱えないという状況であった 2000 年代初頭から、「自然会話データ」を必要とする研究者間で少しでも多くのデータや情報を共有することを最大の目的として構築・公開を始めたのが、本コーパスの前身にあたる諸コーパスである（宇佐美, 2020ab 参照）。有志から自然会話データを提供してもら

いながら、徐々に拡充してきた「拡充型コーパス」で、最初に目的等に応じて計画して収集する「計画型コーパス」とは異なるので、全体を見ると一見条件が統制されていないように見えるかもしれない。しかし、本コーパスの特徴は、会話データが、1～32のフォルダごとにまとめられており、会話フォルダの中のデータは、条件が統制されてデータ収集されている点である。一つの会話グループの中に収められている会話は、3～48と幅があるが、15～20 会話がまとめられているものが平均的である。これらは、なんらかの目的に応じて条件を統制して収集されているので、会話の分析としては、比較的多いといえる 15～20 会話を扱い、定性的分析と定量的分析双方を行うような研究(『総合的会話分析』(宇佐美, 2019 等)に適していると言える。また、語用論的分析には不可欠である「談話の流れ」が可視化されている中で、「同時発話」、「沈黙」、「笑い」、「引用部」など、他のコーパスにはあまり記録されていない情報を記すのがルールとなっている『基本的な文字化の原則 (BTSJ: Basic Transcription System for Japanese) 2019 年改訂版』(宇佐美, 2020a 所収)によって文字化されたトランスクリプトが収録されている。

より具体的には、話者個人の年齢、性別などの個人の社会的属性だけでなく、話者同士の関係(上下関係や、教師と学生等)や面識の度合い(初対面、既知(知り合い)、既知(友人)等)などの情報が統制されて収集されているため、話者同士の関係に応じた「相互作用」の特徴が分析できる。同時発話や沈黙、笑い、引用部など、「語用論的分析」に必須の情報が文字化資料に付与されているため、特定語を検索するという形だけではなく、話者同士のやりとりが、それらの情報を考慮に入れながら「談話の流れ」とともに分析できる。「コア会話」は、主に、初対面会話と友人の会話の比較ができるようになっており、296 会話(母語場面 211 会話、接触場面 85 会話)があり、自然会話データとしては、最大規模の数のデータの比較もできるようになっている。それ以外の多様な場面のデータを含む「非コア会話」には、論文指導場面、討論、依頼、断り会話などに加えて、大学生同士の「ゼミ合宿の食事についての話」や「学校給食についての話」、また「バーベキューの食材についての話」などの大学のキャンパスライフに関する話や、小学生同士の会話、小学生と大人の会話、九州方言の出現を分析するための会話データなど、多様な場面と話題を含む会話が追加され、合計 226 会話が収録されている。

## 2.2 『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』との連動について

2 つめの特徴は、『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』と連動しているので、定量的分析が行いやすくなっている。これらのツールと合わせて活用することができることから、これまであまり定量的分析を行ったことがなかったような人も、データのグローバルな傾向を掴んだ上で、ローカルな分析を行うことができ(宇佐美, 2006)、どちらかというと定性的分析が多い語用論研究の幅を広げることに貢献する研究を促進することができる。ひいては、これまでの定性的分析による語用論研究によって明

らかにされた知見の信頼性を高め、分野全体の質を高めることにもつながることが期待できる。

## 2.3 「自然会話リソースバンク NCRB (National Conversation Resource Bank) との連携について

この『BTSJ 日本語 1000 人自然会話コーパス (完成版・映像付き)』には、100 会話以上の「映像データ」が含まれるため、話者の表情や身体的動きなどの観察もできるようになり、これまで以上に多角的でマルチモーダルな分析が可能になる。また、共同構築型多機能データベースのプラットフォームである NCRB に搭載されるため、ユーザー登録をすれば、NCRB 上の「研究部門」に入ることができ、音声データと文字化資料については、研究用としてダウンロードすることができるようになる。また、NCRB 上の「教材部門」には、「教材作成支援機能」が搭載されているため、映像データは、「自然会話を素材とする教材」の素材とすることもできる。そのため、NCRB 上で、教材用として、解説の入力や Q&A の作成ができるようになる。ここでは、NCRB の機能についても簡単に紹介しておく。

### 2.3.1 研究のためのコーパスデータの利用

①NCRB のトップページから「自然会話データを使った研究」の部分に入り、「会話」というタブをクリックすると、BTSJ コーパスとして、522 会話分のデータが一覧の形で表示される。(今後は、登録メンバーが独自に収集した会話も、管理者の審査を経て、随時、各自で追加できる。)「会話一覧」に表示されている「会話名」をクリックすると、各データの音声データや動画データ、および、それに対応する文字化スクリプトの内容を確認、視聴することができる。また、キーワードの検索機能を使えば、利用者の研究テーマや興味に応じた内容に関係するデータを抽出することができる。さらに、研究のために個別の自然会話データの音声、及び文字化スクリプトをダウンロードして、利用することができる (図 1, 2 を参照)。



図 1 BTSJ コーパスの「会話一覧」画面



図 2 「会話トランスクリプト」画面

②利用者は、BTSJ コーパスの 522 会話全てのデータをダウンロードしなくても、NCRB 上に搭載されている BTSJ コーパスから、各自の研究に必要な会話データだけを集めて、各自が選択したデータの「会話グループ」を作成することができる（名前をつけて保存が可能）。さらに、その「会話グループ」のデータを一括ダウンロードして利用することができる（図 3、4 を参照）。



図 3 会話情報表示画面



図 4 会話グループ作成画面

### 2.3.2. 映像データを利用した「自然会話を素材とする教材」の作成

『BTSJ 日本語 1000 人自然会話コーパス（完成版・映像付き）』のデータを基に研究を行い、得られた知見は、日本語教育の現場において、「自然会話を素材とする教材」の作成にも活用することができる。さらに、NCRB の「自然会話を素材とする教材」の入り口を入ると、「教材作成支援機能」が搭載されているため、動画データを利用して「自然会話を素材とする教材」を作成することもできる。この教材の特徴は、「自然会話」とその「トランスクリプト」をメインとしているところで、そのトランスクリプト上に、主に、フィラーや会話のストラテジーなど、通常の教科書では、あまり説明されることのない要素や現象の解説を入力していく。ここに、BTSJ コーパスも含めて、自然会話を分析した結果得られる知見が有効になってくる。

自然会話教材を作成するためには、まず、自然会話の動画データが必要となる。教材作成の利用者は、自身が撮影した自然会話の動画データを NCRB にアップロードするか、または、他の利用者がアップロードした動画データ（共同編集可となっているもの）を利用することができる。次に、アップロードした自然会話の動画データの文字化資料（トランスクリプト）を作成し、NCRB 上で教材作成の準備を整える。BTSJ コーパスの映像データ付き会話を選択すれば、トランスクリプトは、既にあるので、アップロードすればよい。

NCRB の「自然会話を素材とする教材」には、「自然会話トランスクリプト教材」と「Q&A 教材」の 2 種類ある。

「自然会話トランスクリプト教材」は、NCRB の「教材作成支援機能」を利用し、教材作成

画面に表示されるトランスクリプトの一行一行に、必要に応じて、「内容」「表現」「会話ストラテジー」「ポライトネス」「文化」の5つの観点から該当する学習項目の解説を記入することができる（図5を参照）。

各会話の動画の下部にある「Q&A 教材」には、主に、学習者の理解度（インプット）を確認するために、Step1～3の3つのレベルに応じた設問作成機能が搭載されている。それぞれのStepで「選択問題」と「記述問題」が作成できる（図6を参照）。



図5 各発話の解説を記入する画面

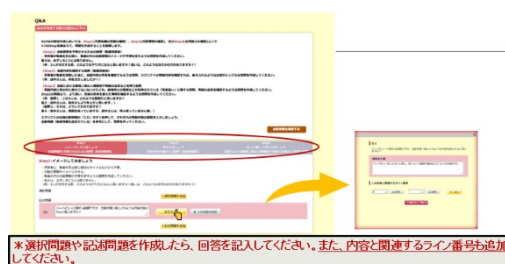


図6 理解度を確認するQ&Aの作成画面

## 2.4 オープンサイエンスの理念を踏まえた21世紀型「共同構築型コーパス」としての新たな展開

以上、『BTSJ 日本語 1000 人自然会話コーパス（完成版・映像付き）』の特徴を3点、まとめてきた。最後の、そして最も重要な本コーパスの特徴が、このコーパスが、21世紀型「共同構築型コーパス」の展開を企図していることである。「1000 人日本語自然会話コーパス」の構築を目標としたプロジェクトとしては、1000 人の日本語会話データの採録を達成することで、一応の「完成」とするが、実は、本コーパスは、今後、上記で紹介した NCRB 上で公開された後は、**21 世紀型「共同構築型」コーパス**として、**オープンサイエンスの理念**を踏まえ、関連研究者が一機関を超えた形で協力し、随時、一定の基準をクリアした自然会話データを拡充していき、関連の研究分野に貢献する基盤を提供するものにしていく予定である。21 世紀型の新しいコーパスとして、従来の「構築する側」と「利用する側」という一方向の構図を転換し、関連する研究者がデータを提供してコーパス構築に貢献するとともに、他の研究者のデータを活用するという双方向の流れを作り、関連する研究者、研究分野全体の発展を企図する。また、そのような形でコーパスを拡充していき、分析して明らかにした知見は、研究者だけでなく、広く社会一般にも何らかの形で還元していこうとするものである。そういう意味では、本プロジェクトは、今、まさに始まったばかりとも言える。

## 3. 『BTSJ 日本語 1000 人自然会話コーパス（完成版・映像付き）』の活用法

以上、BTSJ コーパス独自の様々な特徴をまとめてきた。その多様な機能を理解すれば、その「活用法」については、各自、既に、様々な目的に応じて、多様に膨らんできているの

ではないだろうか。これらの多様な機能を生かしていかに「活用」していくのか、ここからは、各研究者の目的、アイデア次第である。紙幅も尽きてきたので、開発者としては、以下の点を列挙するに留めておきたいと思う。

- ①定量的分析と定性的分析の双方を生かした「語用論的研究」
- ②未だ確立されていないこれまでの言語学の枠に捉われない、人間のやりとりのデータに基づく体系的な「話し言葉の文法」の構築
- ③「対話システム研究」などに取り入れやすいように、ある程度数値化して表せるような、「談話の流れ」や「文脈」を十分考慮した「人間の会話のやりとり」や「話題展開パターン」のモデルの構築
- ④「自然会話を素材とする教材」の作成、自然会話の分析から得られた知見を活かした解説を充実させる経験を踏まえた上での「第二言語習得研究」

#### 4. おわりに

以上、『BTSJ 日本語 1000 人自然会話コーパス（完成版・映像付き）』の特徴と活用法のエッセンスをまとめた。発表では、視聴者からの質問に答える形で、より具体的、且つ、多角的にこのコーパスの特徴、及び、様々な観点から考えられる「活用法」について解説するとともに、フロアの皆さんと活発に意見交換したい。

#### 参考文献

- 宇佐美まゆみ(2006)「談話研究におけるローカル分析とグローバル分析の意義」『言語情報学研究報告 13 自然会話分析への言語社会心理学的アプローチ』:229-243.
- 宇佐美まゆみ (2019)「『総合的会話分析』に基づく研究—『BTSJ 日本語自然会話コーパス』と『自然会話リソースバンク (NCRB)』との連携に触れながら—」,『ヨーロッパ日本語教育』第 23 号, 206-221.
- 宇佐美まゆみ編 (2020a)『自然会話分析への語用論的アプローチ—BTSJ コーパスを利用して—』 ひつじ書房.
- 宇佐美まゆみ編 (2020b)『日本語の自然会話分析 BTSJ コーパスから見たコミュニケーションの解明』 くろしお出版.

【付記】 本研究は、以下のプロジェクトの支援を受けている。

- ・国立国語研究所の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（リーダー：宇佐美まゆみ）
- ・JSPS 科研費基盤研究 A 18H03581「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）
- ・JSPS 科研費挑戦的研究(萌芽)-18K18685「コミュニケーション能力を高める自然会話教材の高度共有化—共同体の構築に向けて—」（研究代表者：宇佐美まゆみ）