『BTSJ日本語自然会話コーパス(2018年版)』構築の趣旨と特徴

『BTSJ 日本語自然会話コーパス（2018 年版）』構築の趣旨と特徴

宇佐美まゆみ，山崎誠
人間文化研究機構国立国語研究所
{usamima, yamazaki}@ninjal.ac.jp

1 はじめに

国立国語研究所では、「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）の研究成果として、『BTSJ 日本語自然会話コーパス（トランスクリプト・音声）2017 年先行リリース版』を 12 月に内部公開し、2018 年 4 月の一般公開に備えている 。本発表では、現在公開されている音声・文字化資料付きの「自然会話」コーパスとしては国内外を含めて最大規模の本コーパスの構築の趣旨と特徴を 2018 年 4 月の一般公開に先駆けて紹介し、公開後のスムースな活用を促したい。

2 『BTSJ 日本語自然会話コーパス（2018 年版）』の構築・公開の趣旨

近年、自然会話分析が数多く行われるようになり、話し言葉コーパスもいくつか公開されるようになってはきたが、音声学的な分析や、形態素解析、構文の分析のためではなく、人間の相互作用としての「言語運用」の語用論的分析に適した形で文字化され、蓄積された「話し言葉のコーパス」は、未だほとんどないのが現状である。その一つの理由として、自然会話をデータとして用いる研究では、会話の収集、文字化といった基礎的作業に多大な時間と労力を要するということがある。そのため、このような研究を効率的に進めていくには、自然会話データを研究者間で共有化することが必須である。しかし、言語処理やコーパス言語学的アプローチ以外の言語研究においては、比較的少数のデータを各研究者が収集し質的に分析するケースが多く、プライバシー保護の観点からも、コーパス化や共有化は進んでいないのが現状である。

そういう状況の中、宇佐美研究室[1]では、多様な場面・言語（日本語、韓国語、中国語、英語など）の自然会話データを収集し、膨大な時間と労力を投入して『BTS（Basic Transcription System）による多言語話し言葉コーパス』の構築・公開・拡張に取り組んできた。2017 年の段階では、『BTSJ による日本語話し言葉コーパス（トランスクリプト・音声）2011 年改訂版』294 会話、4000 分 31 秒（約 66 時間）を公開し、申込者に無料で配布している[2]。このコーパスに、さらに 39 会話 753 分 15 秒（約 12 時間）のトランスクリプトと音声データを追加し、また、既存のトランスクリプトに、音声データ 24 会話 401 分 5 秒（約 6 時間 40 分）を追加し、それらを全てまとめたものが『BTSJ 日本語自然会話コーパス（トランスクリプト・音声）2018 年版』である。

本コーパスには、333 会話、総時間 4746 分 44 秒（約 79 時間）の会話が収録されており、そのうち音声付きデータは 203 会話、2402 分 42 秒（約 40 時間）である。整備にあたっては、記号などの表記を「基本的な文字化の原則（BTSJ: Basic Transcription System for Japanese） 2015 年改訂版」に改めてある。

本コーパス構築の趣旨は、「相互行為としての会話」の対人コミュニケーション論、語用論的分析に適したコーパスを構築することである。そのために重視した点は、以下の 3 点である。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、話題などを統制したデータ群を収録する。②発話の重なりや沈黙など、語用論的分析に不可欠な情報を記して細やかな定性的分析を可能にするとともに、各研究者が独自の観点から分析項目のコーディングや集計などの定量的分析が行いやすい文字化のルールである「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の形で提供する。③「人間

[1] 2007 年～2015 年は、東京外国語大学、2016 年以降は、国立国語研究所で構築を行っている。

[2] 以下の URL からオンライン申請が可能である。https://ninjal-usamilab.info/btsj_corpus/

の相互作用としての会話分析」は、「会話自体」の分析のみならず、「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。

このように、当コーパスに収録された会話は、グループごとに、収集の目的や、会話の条件が統制されているため、グループごとの目的・条件を確認し、研究目的に応じて、話者の属性（年齢，性別等）や対話相手との関係など、話者の話し方に大きな影響を与える社会的要因を考慮に入れた分析が可能である。これが、本コーパスの最大の特徴である。そういう意味で、「人間の相互作用の分析に適したコーパス」であると位置づけている。本コーパスの公開の最大の目的は、未だ質的分析に留まっている「言語運用」に重きをおいた「語用論的研究」の妥当性や信頼性を高めるためにより多くの条件統制されたデータで計量的に検証できるようにし、人間の相互作用、言語運用に重きをおく「語用論的研究」の幅を広げるとともに、言語処理研究者にも、大量データの形態素解析などのような機械的な言語形式の分析だけではなく、**話者間の上下、親疎関係など、実際の言語運用と人間関係の構築に極めて重要な情報**や**文レベルを超えた談話の流れ**を十分に考慮した分析を促すことによって、自然会話をデータとする言語運用、人間の相互作用の研究の発展を促進することである。

3 『BTSJ 日本語自然会話コーパス（2018年版）』の総合的特徴

本コーパスでは、会話参加者の年齢、性別、話題などが統制された形でデータが収集されており、様々な観点から比較・対照研究ができるようになっている。また、BTSJ の背景理論となる言語社会心理学、及びその方法論である「総合的会話分析」（宇佐美、2008；2013；2015）では、会話自体の分析のみならず、データの収集法、被験者の属性調査など、「録音された会話」以外の部分の分析も、人間の相互作用としての「会話分析」のために、極めて重要だと捉えているため、条件が統制された各会話グループのデータ収集計画や話者の年齢・性別・属性等のデータベースも含まれている。条件が統制されていれば、大規模データにおけるサブ・グループの結果と同様の傾向を示すことも、ある程度確認されている（宇佐美・中俣、2013）。

以下の表 1 に、本コーパスの概要を、表 2 に各会話グループの条件、属性ごとの話者数を示す。

表 1『BTSJ 日本語自然会話コーパス（トランスクリプト・音声）2018 年版』の会話データの概要

本コーパスにおける会話の通し番号	会話グループ番号	登録者の会話グループ名	各グループのデータの特徴	各グループ内のデータ数	各グループの総会話時間
1-19	1	親しい同性友人同士雑談（男性、女性）	同性の友人同士の会話	19 会話	444 分 24 秒
20-42	2	初対面及び友人同士雑談（女性）	女性の、親しい友人同士と初対面の会話	23 会話	482 分 5 秒
43-52	3	論文指導（日本人教師男女、日本人学生男女）	教師と学生の面談の会話	10 会話	311 分
53-91	4	女性同士の断りの電話会話（対先輩、対同級生、対後輩）【音声付】	ある学生（女性）をベースに、電話で、先輩・同輩・後輩に依頼をする会話	39 会話	78 分 31 秒
92-111	5	同性同士の依頼を含む電話会話（男性、女性）	同性の友人同士の会話	20 会話	53 分 02 秒
112-116	6	友人同士雑談（女性）	女性の友人同士の会話	5 会話	80 分 41 秒
117-120	7	OPI インタビュー(テスター男性、受験者女性)【音声付】	OPI インタビュー形式に基づく、フランス語母語話者の縦断的データ	4 会話	41 分 25 分
121-129	8	韓国人学習者（中級男性、中級女性）と日本人の初対面同性同士雑談	韓国人日本語学習者と日本人の初対面同性同士の雑談	9 会話	249 分
130-141	9	台湾人学習者ベース（上級男性、上級女性）と日本人（年上、同等）の初対面同性同士雑談	台湾人日本語学習者の接触場面データ	12 会話	234 分 20 秒

142-151	10	台湾人学習者(上級)と日本人の友人同士雑談(女性)【音声付】	台湾人日本語学習者の接触場面データ	10 会話	173 分 30 秒
152-160	11	日本人女性ベース初対面同性同士雑談(日本人、台湾人中級学習者、台湾人超級学習者)【音声付】	20 代前半の日本人女性(学生)が、対同世代の日本人女性、対日本語中級話者、対日本語超級話者と 3 通りの会話を行っている	9 会話	159 分 48 秒
161-172	12	日本人女性ベース初対面同性同士雑談(日本人、ベトナム人初級学習者、韓国人初級学習者、中国人上級学習者)	20 代前半の日本人女性(学生)が、対同世代の日本人女性、対日本語初級話者、対日本語上級話者と 3 通りの会話を行っている	12 会話	120 分 11 秒
173-190	13	男性ベース初対面雑談(同性目上、異性目上、同性同等、異性同等、同性目下、異性目下)【音声付】	35 歳男性が、年上(45 歳)・同等(35 歳)・年下(25 歳)の話者(男/女)と 6 通りの会話を行っている	18 会話	299 分 34 秒
191-206	14	初対面男女、同性同士雑談(同等、目上) 【音声付】	20 代前半大学生・大学院生、初対面の雑談	16 会話	268 分 55 秒
207-209	15	友人同士雑談(女性)	20 代女性学生、親しい友人同士の雑談	3 会話	63 分 37 秒
210-257	16	友人同士雑談及び討論(同性、異性)【音声付】	日本語母語話者、10 代後半から 20 代前半の大学生、ベース話者男女各 6 名が、「同性/異性」の友人と、「雑談/討論」という 4 通りの会話を行っている	48 会話	750 分 24 秒
258-262	17	友人同士討論(異性)	20 代-30 代学生、友人同士の討論	5 会話	88 分 16 秒
263-266	18	初対面討論(女性同士)	20 代女性、大学生・大学院生、初対面の討論	4 会話	44 分 33 秒
267-274	19	友人同士誘い(女性)	20 代大学生友人同士。話者の一方が協力者である。協力者が「気軽に行うこと」を誘うように依頼した	8 会話	172 分 53 秒
275-286	20	日本人女性と日本人、台湾人上級学習者、中国人上級学習者の初対面同性同士雑談【音声付】	日本語母語話者同士の会話と、日本語母語話者と日本語学習者の会話	12 会話	186 分 20 秒
287-318	21	女性同士の謝罪のロールプレイ会話(負担の重い場面、負担の軽い場面)【音声付】	2 人の話者が、負担度の軽い場合と重い場合の 2 つの謝罪場面についてロールプレイを行っている	32 会話	76 分 19 秒
319-328	22	中国人女性学習者(初級、上級)と日本人友人同性同士雑談【音声付】	中国人日本語学習者(初級 5 名、上級 5 名)と日本語母語話者の女性友人同士の雑談	10 会話	262 分 44 秒
329-333	23	初対面及び友人同士雑談(女性同士)【音声付】	20 代前半、女子大学生同士の雑談(初対面 2 組、友人 3 組)	5 会話	106 分
計				333 会話	4747 分 32 秒 (約 79 時間)

表 2 会話グループの条件、属性ごとの話者数

会話グループ	母語話者・初対面		母語話者・友人		非母語話者・初対面		非母語話者・友人	
母語場面/接触場面	母語場面		母語場面		接触場面		接触場面	
話者の関係	初対面		友人		初対面		友人	
性別組み合わせ	女性同士	72	女性同士	214	女性同士	34	女性同士	20
	男性同士	34	男性同士	64	男性同士	9	男性同士	0
	男女	18	男女	58	男女	0	男女	0
話者総数		124		336		43		20
年齢	20 代	94	10～20 代	326	20 代	32	10～20 代	20
	30 代	24	20～30 代	10	30 代	2	30 代	0

	40 代	6			不明	9		
非母語話者の出身					韓国	11	台湾	10
					台湾	17	中国	10
					中国	7		
					ベトナム	2		
非母語話者の日本語レベル					超級	3	超級	0
					上級	24	上級	15
					中級	12	中級	0
					初級	4	初級	5

4 『BTSJ 日本語自然会話コーパス（2018 年版）』の基本情報

表 3 に本コーパスの基本情報を示す。形態素解析は、MeCab 0.966 及び現代話し言葉 UniDic（unidic-csj-2.2.0）を使用した。語数は短単位による数字である。なお、表中の語数の集計に際しては、解析結果において品詞が「補助記号」となったもの（主に句読点など）は含めていないが、品詞が「記号」の場合は数値に含めている。

表 3 『BTSJ 日本語自然会話コーパス（トランスクリプト・音声）2018 年版』の語数等の基本情報

会話数	333 会話
延べ語数(Token)	928,102 語
異なり語数(Type)	13,498 語
Type/Token 比	0.0145
発話文数	104,489 文
1 会話あたりの語数	2,787.09 語
1 文あたりの語数	8.882 語
総時間	284,784 秒 （79 時間 6 分 24 秒）
1 文あたりの時間	2.725 秒
話者数（異なり）	224 人

以下の図 1 に、トランスクリプトのイメージを示す。

図1 BTSJ システムセット

5 まとめと今後の予定

本コーパスは 2018 年 4 月に一般公開を予定している。今回は、本コーパスの全体的特徴を示すために、また、紙幅の制約もあり、基礎統計量等を示すに留めたが、実際の分析をする際は、言語形式の頻度を比較するような分析だけではなく、各研究者がそれぞれの観点からコーディングを行った上で、量的・質的両方の分析を行うことを推奨したい。また、言語処理の分野においても、人間の相互作用や人間関係の分析につながるような言語運用の研究もさかんになることを期待したい。

【参考文献】

宇佐美まゆみ（1999）「談話の定量的分析 -言語社会心理学的アプローチ-」『日本語学』18(11)、明治書院：40-56.

宇佐美まゆみ（2008）「相互作用と学習－ディスコース・ポライトネス理論の観点から」西原鈴子・西郡仁朗編『講座社会言語科学 第 4 巻 教育・学習』、ひつじ書房：150-181.

宇佐美まゆみ（2013）「会話データの作成・分析―「総合的会話分析」と「基本的な文字化の原則（Basic Transcription System for Japanese:BTSJ）」」『日本語学』32(14)、明治書院：132-147.

宇佐美まゆみ・中俣尚己(2013)「『BTSJ による日本語話し言葉コーパス（トランスクリプト・音声）2011 年版』の設計と特性について」『第 3 回 コーパス日本語学ワークショップ予稿集』、217-228.

宇佐美まゆみ（2015）「『総合的会話分析』の趣旨と方法－量的分析と質的分析の必然的融合－」、日本語教育 162 号、34-49.

山崎誠（2017）「レジスター・位相の違いによる会話文の語彙的多様性」、『言語資源活用ワークショップ 2017 発表論文集』、278-289.