『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

宇佐美 まゆみ（東京外国語大学大学院総合国際学研究院）†
中俣 尚己（京都教育大学教育学部）‡

Design and Characteristics of the "Corpus of Spoken Japanese by BTSJ (Transcription and Audio Recordings) ver.2011"

Mayumi Usami (Graduate School of Tokyo University of Foreign Studies, Institute of Global Studies)
Naoki Nakamata (Kyoto University of Education)

１．はじめに

近年、コーパス日本語学が盛んになりつつあるが、その多くは書き言葉のコーパスに関するものであり、「話し言葉コーパス」に基づくものは多いとは言えない。分析の観点も、書き言葉の特性を考えると当然かもしれないが、形態素解析や語彙や構文の分析、コロケーション研究などが中心で、語用論的分析は未だ手つかずの状態である。一方、「話し言葉のコーパス」も増えつつはあるが、人間の相互作用としての「自然会話（事前の計画がないやりとり）」を編んだコーパスは、ほとんどないといっても過言ではない。日本語学習者の口頭能力試験を集めた学習者コーパスなどはいくつかあるが、これらは口頭能力試験という特殊な状況における相互作用であり、分析の観点も、未だ文法項目の習得などに焦点を当てたものが多い。「会話分析」としては、エスノメソドロジーに端を発する CA（Conversation Analysis）が盛んであるが、基本的に、CA は、対人コミュニケーションの理論化や一般化を目的とはしていないこともあり、その「文字化システム」は、「定性的分析」には適しているかもしれないが、「定量的分析」には適しているとは言えない。昨今公開されている「話し言葉のコーパス」も、語用論的分析に適した「文字化システム」に基づくものはほとんどない。話し言葉コーパスに基づく分析は、講演などのストレート・トークやナラティブ・データに基づいた音声学的な分析などが緒についたところであると言ってもよいだろう。すなわち、人間の相互作用の分析を企図し、会話の定性的分析に加えて、定量的な分析も可能にする形で文字化し蓄積された「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。その理由の一つに、話し言葉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をはじめ、その後の分析対象のコーディングなどにも膨大な時間と労力を要するということがある。そのため、会話や話し言葉の対人コミュニケーション論的・語用論的分析を、より効率的に進めていくためには、研究者間で自然会話データを共有していくことが不可欠である。また、そのためには、発話の重なりや沈黙などの語用論的分析に必須の情報を記述し、且つ、定量的分析にも適する文字化システムによって蓄積された「話し言葉コーパス」が必須である。このような認識に基づいて、筆者とその研究協力者らは、ここ 15 年来、あくまで人間の相互作用としての「言語の運用」に焦点を当て、対人コミュニケーション論、語用論の観点から「会話の分析」を行い、定量的分析ができる形で文字化したデータを蓄積し、一般公開も行ってきた。それらを改訂し、改めてまとめ直したのが、『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』（以降「BTSJ 話し言葉コーパス」と略記）である。本稿では、その開発・設計の趣旨、及び、その特性と活用方法を簡単にまとめる。

† usamima@tufs.ac.jp ‡ nakamata@kyokyo-u.ac.jp

２.『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』の設計の趣旨と特性

本節では、BTSJ 話し言葉コーパスの設計の趣旨と特性を簡単にまとめる。

２．１　BTSJ 話し言葉コーパス設計の趣旨

「１．はじめに」でも述べたように、本コーパス設計の趣旨は、「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、分析項目のコーディングや集計などの定量的分析も行いやすい「基本的な文字化の原則」である 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、その他の属性をまとめたエクセルファイルも収録する。

２．２　BTSJ 話し言葉コーパスの概要と特徴

『BTSJ による日本語話し言葉コーパス』は 294 の相互作用的会話からなる[1]。会話の総時間は 67 時間 21 分 39 秒、総語数は、789,190 語[2]である。すべての会話は、発話の重なりや沈黙、割り込みなどの語用論的分析に必須の情報を記述するための原則である「基本的な文字化の原則（Basic Transcription System for Japanese：BTSJ）2011 年版」に基づくトランスクリプトの形になっており、約 30%（20 時間分）の会話には、プライバシー保護処理をした「音声資料」がトランスクリプトとともに提供されている。BTSJ トランスクリプトは、多くの人が活用しやすいことを考え、エクセル形式で保存されている。利用する研究者各自が、「発話内容」の右側に「コーディング」の列を追加して分析したい項目をコーディングすれば、エクセルの機能で話者ごとにソートして話者の特徴を概観したり、コーディング項目の頻度の集計などを行うことができるが、2007 年に、エクセルに専用のマクロ機能を搭載して「BTSJ 入力支援・自動集計システムセット」を開発し、対となる記号の自動入力やエラーチェック機能等の「入力支援機能」を付与し、コーディング項目の基本的記述統計を自動集計して表の形で自動表示できるようにした。さらに、2011 年には、同じルールでコーディングした複数の会話ファイルの分析項目の頻度や割合の合計、平均、標準偏差などの自動集計も可能にした。このシステムセットは、現在のところ、「BTSJ 活用方法講習会」[3]（宇佐美 2012）の受講者に無償で配布している。また、テキストファイルに変換して利用することもできる。本コーパスは、事前の計画や準備のない自然会話を中心とするコーパスであるが、一部、電話会話やロールプレイ等も収録されており、日本語母語話者の会話のみならず、接触場面（日本語母語話者と日本語非母語話者）の会話も豊富である。初対面、友人同士、話者の年齢に上下のある会話、同年齢同士の会話、同性同士の会話、異性との会話、教師と学生の面談会話等々、様々な種類の会話が、話者の社会的属性や場面等の諸条件を統制して収集され、収録されている。そのため、話者の社会的属性や話者同士の関係、場面に応じた話し方の特徴や違いを、様々な角度から比較・検討することが可能である。この点が、BTSJ 話し言葉コーパスの最大の長所であり、特徴である。

---

[1] 予稿集では、改訂中の 1 会話を除いた数値を提示したが、本稿では、コーパスに収録されている 294 会話すべてを含めて算出した数値を提示する。これ以降の表 1 などの数値についても同様である。

[2] Mecab+UniDic による。句読点等を除く実質的発話部である。

[3] これまでのところ不定期に、東京、広島、京都、九州、ベルリン、ロンドンで開催している。問い合わせ：言語社会心理学研究会事務局：btsjworkshop@gmail.com

２．３　BTSJ（Basic Transcription System for Japanese）の基本原則と形式

すべてのトランスクリプトは、BTSJ によって記述されており、xlsx 形式のエクセルファイルで提供される。BTSJ によるトランスクリプトの一例を以下の図１に示す。

上部には「会話グループ名」、「会話記号（ファイル名に対応）」、「話者記号の凡例」、「会話番号」、「時間」、「１会話における話者数」の６つの情報が記載されている。その下に発話内容（トランスクリプト）が記される。左には「ライン番号」、「発話文番号」、「発話文終了」、「話者」を記す。

BTSJ では、「発話文」の定義は、「会話という相互作用の中における文」とし、以下のように認定する。基本的に、ひとりの話者による「文」を成していると捉えられる発話を「1 発話文」とする。しかし、自然会話では、いわゆる「1 語文」や、述部が省略されているもの、あるいは、最後まで言い切られない「中途終了型発話」など、構造的に「文」が完結していない発話もある。そのような場合は、話者交替や間などを考慮した上で「1 発話文」であるか否かを判断する。つまり、「発話文」の認定には、「話者交替」、「間」という 2 つの要素が重要になる。そのため、途中で相手の発話が入って話者が一旦交替したため改行され、複数のラインに渡っている発話も、同一話者によって発せられた「１文」を成していると捉えられるものは、複数のラインにまたがる発話をまとめて「１発話文」とする。そして、図１の「発話文番号」の列における「3-1」、「3-2」のように、異なるラインにまたがっていても同じ発話文であることがわかるように同じ番号をつけ、その後に「-」をつけて発話された順を記す。また、完結していないほうの発話には、「発話文終了」の欄に「/」を記す。１会話の「発話文数」は、「発話文番号」が示すとともに、左から３行目の「発話文終了」の列が、発話文が完結していることを表す「*」となっているものを数えてもわかるようになっている。また、「発話内容」の列における「。」も、BTSJ のルールでは発話文の完結を意味するため、質問発話で文末に「?」があっても、文が完結している場合は、「?。」と、必ず、最後に「。」をつける。そのため、「*」と「。」の数が同じになることを利用して、通常のエクセルでも「*」と「。」を数えることによって、発話文数の検算もできる。

会話グループ名：台湾人学習者（上級）と日本人の友人の雑談	会話記号：TF01-JF01	記号凡例： TF：Taiwanese Female JF：Japanese Female
会話番号：142	時間：0分0秒-20分40秒(終)	1会話における話者の数：2

ライン番号	発話文番号	発話文終了	話者	発　話　内　容
1	1	*	TF01	####しゃべってください。
2	2	*	JF01	えっ、何をしゃべるんですか<笑い>。
3	3-1	/	TF01	《沈黙2秒》ね、しゃべらないと,,
4	4	*	JF01	<笑いながら>しゃべれないよ。
5	3-2	*	TF01	また私ずっとしゃべってる感じだ(<笑い>)と思いました。
6	5	*	TF01	あっ[[。
7	6	*	JF01	]]あっ、お菓子食べる<時に…>{(}。
8	7	*	TF01	<そう>(}。
9	8	*	JF01	かな。

図１　BTSJ によるトランスクリプトの例

発話内容には種々の記号を用いて、相互作用に関する情報が付与されている。図１には「沈黙」「笑い」「発話の重なり」「さえぎり」等の情報が記載されている。その他にも「引用部」「イントネーション」「ラッチング」「言い淀み」「文脈情報」などの情報が付与されている（記号の意味など、BTSJ に関する詳細は、宇佐美(2011)を参照）。ただし、各研究者が自身のデータを BTSJ で文字化する場合は、研究目的に応じて、BTSJ で定められた記号を変更しない限りにおいて、独自の記号を追加して特定の現象の記述をより詳細にしたり、逆に、小声のあいづちは文字化しない等の原則を設けて簡略化することも可能である。

２．４　『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（2012 年改訂版）』について

BTSJ は、あくまで「文字化のルール」である。そして、『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（2012 年改訂版）』は、BTSJ による「文字化」にかかる時間と労力を軽減するための「文字化入力支援機能」と、BTSJ で記されたトランスクリプトにコーディングを行った項目の基本的な記述統計に必要な情報を算出する「自動集計機能」を搭載したシステムセットである。利用者の利便性や汎用性を考えて、Microsoft Excel のマクロ機能を利用して作成されており、「BTSJ 入力支援・自動集計システム(.xlt)」、「BTSJ 複数ファイル自動集計システム(.xls)」の 2 つのファイルから成っている。現在は、日本語版 Windows の Excel 2003、2007、2010 に対応している。（ただし、英語版 Windows でも、Excel 上で日本語を表示できる環境であれば、問題なく使える。また、Mac の場合は、windows をインストールするか、シトリックス http://www.apple.com/jp/business/profiles/citrix/ などの仮想デスクトップを導入する必要がある。）

３．『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』の定量的な基本情報

本節では、『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 年版』の定量的な基本情報を、『現代日本語書き言葉均衡コーパス』（以下 BCCWJ と略記）と比較する形で示す。

３．１　基本情報

本節では、「BTSJ 話し言葉コーパス」に形態素解析を施した結果を示す[4]。まず、表１に、「総語数」、「異なり語数」などの本コーパスの基本情報を示す。

表１　『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』の基本情報

会話数	294 会話
総語数	789,190 語
異なり語数	12,079 語
TTR（異なり語数／総語数）	1.530%
Guiraud 値（異なり語数／√総語数）	13.596
発話文数	91,256 文
１文あたりの語数（総語数／発話文数）	8.648 語
総時間	242,499 秒（67 時間 21 分 39 秒）
１文あたりの時間数（総時間／発話文数）	2.657 秒

注）なお、上記の語数には、句読点など、UniDic において「補助記号」に分類されるものは含まない。また、「記号」は、人名などが記号で表されることもあるため（例：F さん）、数値に含めている。

３．２　動詞の高頻度語

「基本語彙」の選定は外国語学習の分野において極めて重要である。本節では、まず、BTSJ 話し言葉コーパスにどのような動詞が多く見られたかを算出し、BCCWJ と比較する。

次頁の表２に、BTSJ 話し言葉コーパスにおける高頻度の動詞、上位 20 と、BCCWJ における高頻度の動詞、上位 20 を、それぞれ 1 万語あたりに換算し頻度とともに示す。太字ゴシックの語は当該コーパスでのみ上位 20 以内に入った語である。表２を見ると、上位 20 語のうち太字ゴシックの語を除く 16 語までが、両コーパスに共通していることがわかる。

[4] エクセルファイルのコーパスを csv 形式に変換後、発話内容の部分だけを取り出し、BTSJ 特有の記号を除去した後、茶まめ(UniDic+mecab)を用いる形で形態素解析を行った。

基本語彙の選定は、これまでも種々の立場から行われているが、中でも最も数が絞られているのは、国立国語研究所の『電子計算機による新聞の語彙調査』をもとに、林(1975)が「文句なしの基本語彙」とした 545 語であろう。その中に動詞は 73 語ある。BTSJ 話し言葉コーパスの高頻度語 20 の中では、「違う」と「書く」を除く 18 語がこの 73 語に含まれている。

また、BTSJ 話し言葉コーパスのみで上位 20 に入った「違う」「聞く」「書く」「取る」の 4 語は、BCCWJ でも、40 位以内に入っている。このことを考えると、これら 4 語は、書き言葉においても基本語に相当すると言ってもよいだろう。つまり、動詞の高頻度語は、コーパスの規模の大小、話し言葉、書き言葉の違いにかかわらず、ほとんど共通しているということと、「基本語彙」（林 1975）との共通性が高いということが明らかになった。

一方、BCCWJ では 12 位に入っている「おる」は、BTSJ 話し言葉コーパスでは、20 位までには入らず、61 例（1 万語あたり 0.77）しかなかったが、実は、BCCWJ の中でも、「国会会議録」に集中的に出現する語であることがわかった。このように、語用論の観点からは、大規模コーパス全体における単なる頻度の比較ではなく、ジャンルごとに分けてみた頻度やコロケーションの分析・考察が重要である。

表２　動詞の高頻度語の比較

順位	BTSJ 話し言葉コーパス	1 万語あたりの頻度	BCCWJ	1 万語あたりの頻度
1	言う	143.10	いる	109.73
2	する	119.69	する	60.51
3	ある	65.66	なる	48.69
4	行く	50.09	ある	47.20
5	思う	48.02	言う	30.43
6	やる	36.14	来る	22.83
7	いる	36.04	思う	20.35
8	なる	32.67	できる	13.29
9	来る	31.37	見る	18.14
10	分かる	23.61	行く	15.36
11	見る	23.12	**しまう**	9.69
12	**違う**	13.56	**おる**	9.60
13	できる	13.29	考える	9.34
14	入る	10.66	**持つ**	8.48
15	出る	10.06	分かる	8.08
16	**聞く**	9.73	出る	8.01
17	**書く**	9.41	やる	7.80
18	知る	8.77	**行う**	5.95
19	考える	7.39	知る	5.79
20	**取る**	7.31	**入る**	5.69

注 1)BCCWJ における「てる」は UniDic では助動詞となっているため、除外した。

注 2)BCCWJ のデータは Ninjal-LWP for BCCWJ Ver.1.10 を使用したため、BCCWJ のうち約 6 千万語分のデータにおける順位である。

３．３　副詞の高頻度語

次に、動詞と同様、BTSJ 話し言葉コーパスにおける副詞の高頻度語を、BCCWJ と比較する形で示す。次頁の表 3 に、BTSJ 話し言葉コーパスにおける高頻度の副詞上位 20 と、BCCWJ における高頻度の動詞上位 20 を、それぞれ 1 万語あたりの頻度とともに示す。

太字ゴシックの語は当該コーパスでのみ上位 20 に入った語である。動詞の結果とは対照的に、BTSJ 話し言葉コーパスの上位 6 語こそ BCCWJ においても上位語になっているが、7 位以下の 14 語のうち 12 語までが、BTSJ 話し言葉コーパスでのみ上位に入った語となっている。同様に、BCCWJ でも上位 20 語のうち 12 語は、BTSJ 話し言葉コーパスでは上位 20 に入っていない。また、先述した林(1975)の基本語彙の中には副詞が 55 語含まれているが、BTSJ 話し言葉コーパスにおける高頻度副詞 20 のうち、この副詞 55 語に含まれているのは、「もう」「やはり」「あまり」「まだ」「もし」「もっと」「いっぱい」「例えば」の 8 語のみであった。林(1975)は、新聞の語彙調査を元にしていることから、副詞の基本語彙は、話し言葉と書き言葉で、かなり異なっていることがわかる。

これらの結果から、副詞の用法は、話し言葉と書き言葉の違いを特徴づける語群の一つであると言えるだろう。

表３　副詞の高頻度語の比較

順位	BTSJ 話し言葉コーパス	1 万語あたりの頻度	BCCWJ	1 万語あたりの頻度
1	そう	167.91	そう	7.62
2	もう	36.21	どう	6.91
3	ちょっと	31.20	もう	5.14
4	こう	27.93	**さらに**	3.77
5	やはり	25.49	やはり	3.24
6	どう	25.22	まだ	3.04
7	**まあ**	22.93	よく	2.87
8	**結構**	16.16	**少し**	2.87
9	**あまり**	12.87	**すぐ**	2.52
10	**多分**	12.16	**まず**	2.52
11	**全然**	11.77	**特に**	2.48
12	まだ	8.52	**まったく**	2.37
13	よく	6.69	ちょっと	2.21
14	**ずっと**	5.12	**すでに**	2.14
15	**色々**	5.08	こう	2.08
16	**なるほど**	4.94	**実際**	2.02
17	**うんうん**	4.80	**ほとんど**	1.85
18	**例えば**	4.75	**最も**	1.74
19	**一番**	4.22	**初めて**	1.73
20	**ちゃんと**	3.78	**もちろん**	1.68

表 3 から、「そう」が双方のコーパスで 1 位であり、「やはり」が BTSJ 話し言葉コーパスで 6 位、BCCWJ で 5 位であることなどから、一見両コーパスで同様の傾向を示しているように見える。しかし順位が同じでも、1 万語あたりの頻度を見ると、話し言葉のほうがかなり多い。また、用例に目を通すと、「そう」は BTSJ 話し言葉コーパスでは、ほとんどが「あ、そうなんだ」「そうそう」「そうですか」のように応答に使われているのに対して、BCCWJ では「母はそう言った」のように具体的な指示内容を持つ用法が多いことがわかった。また、「やはり」の音形に着目すると、BTSJ 話し言葉コーパスでは、「やはり」（4%）、「やっぱり」（61%）、「やっぱ」（35%）であるのに対して、BCCWJ では、「やはり」（68%）、「やっぱり」（28%）、「やっぱ」（3%）となり、話し言葉と書き言葉の違いが顕著に見えてくる。このように、話し言葉と書き言葉の特徴を比較するためには、単なる順位や頻度の比較だけではなく、用例や音形、コロケーションなども考慮に入れた分析が必須である。

４.『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』を用いた語用論的分析

ここまでは、BTSJ 話し言葉コーパスの語彙的特性の概観を定量的観点から示した。しかし、本コーパスは、話者の社会的属性や場面などが統制されて収集されていることが最大の特徴であり、特定の場面やある属性をもつ話者のみを取り出して、その特徴や言語使用の分析を行うことのほうを重視している。この点は、様々のジャンル、媒体、文体等のデータを収録した大規模コーパスに基づいて、データのジャンルや属性の違いをあまり考慮せずに分析している研究が多い従来の「コーパス言語学」においては、あまり重視されていない点である。本節では、語用論的分析の一つとして、話者の属性（母語話者／非母語話者)、話者同士の関係（初対面／友人)、場面（母語場面／接触場面）を統制した形で、それぞれの条件における「異なり語数」と発話文末の「丁寧体率」（丁寧体／総発話文数）を算出することによって、それぞれの状況や話者の属性による語彙数やスピーチレベルの違いを明らかにする。

４．１　話者同士の関係、場面の違う会話における母語話者と非母語話者の「異なり語数」の比較

ここでは、上下関係のない２人の話者の会話において、丁寧体（「です」「ます」）の使用率が、話者の属性（母語話者／非母語話者)、話者同士の関係（初対面／友人)、場面（母語場面／接触場面）によって、どのように異なるかを分析する。そのため、「BTSJ 話し言葉コーパス」の中から、これらの条件に相当する会話を選び出し、「母語場面・初対面」「母語場面・友人」「接触場面・初対面」「接触場面・友人」の４つのグループに分け、それぞれの会話の発話内容のみを取り出し、茶まめ（mecab+UniDic）で形態素解析を行った。母語話者同士の会話である「母語場面・初対面」と「母語場面・友人」は、ファイル内のすべての発話内容を分析対象とし、母語話者と非母語話者の会話である「接触場面・初対面」と「接触場面・友人」では、非母語話者の発話だけを抽出することによって、母語話者と非母語話者という話者の属性による違いを分析した。以下の表４に、各会話グループの条件、属性ごとの話者数を示す。

表４　各会話グループの条件、属性ごとの話者数

グループ名	母語話者・初対面		母語話者・友人		非母語話者[*1]・初対面		非母語話者[*1]・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	24	女性同士	18	女性同士	24	女性同士	10
	男性同士	7	男性同士	3	男性同士	4	男性同士	0
	男女	3	男女	3	男女	0	男女	0
話者総数		66		54		28		10
年齢	20 代	54	20 代[*2]	54	20 代	28	20 代	10
	30 代	12	30 代	0	30 代	0	30 代	0
非母語話者の出身					台湾	24	台湾	10
					中国大陸	4	中国大陸	0
非母語話者の日本語レベル					超級	3	超級	0
					上級	22[*3]	上級	10
					中級	3	中級	0

*1　接触場面の会話における非母語話者を対象としている。　*2　10 代後半の数名を含む。
*3　うち４名は中国大陸出身である。

次に、各グループの語数などの基本情報を表５に示す。

表５　各会話グループの語数

グループ名	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
延べ語数	117,497	103,534	39,386	15,801
異なり語数	4,002	4,391	2,113	1,572
Guiraud 値	11.68	13.65	10.65	12.51

*接触場面における非母語話者の発話のみを分析対象としている。

表５において、語彙の豊富さの指標となる Guiraud 値を見ると、全体的には、母語話者のほうがやや高いが、母語話者の初対面会話よりも、非母語話者の友人場面のほうが、Guiraud 値が高くなっている。また、母語話者、非母語話者ともに、初対面会話より友人との会話のほうが、語彙使用の幅が広いということがわかる。これらの結果から、異なり語数については、母語話者、非母語話者の違いよりも、初対面会話か、友人との会話かという場面による違いのほうが大きいということが明らかになった。これは、初対面会話では、互いの自己紹介のように話題が画一的なものになる傾向があり（宇佐美・嶺田 1995)、用いられる語彙も限られたものになりがちであるということを示していると言えるだろう。

４．２　話者同士の関係、場面の違う会話における母語話者と非母語話者の「丁寧体率」の比較

日本語における対人コミュニケーションにおいては、相手や場面に応じて、丁寧体と普通体がどのように使い分けられるかということは、対人関係調整上、重要な意味を持つ。しかし、非母語話者にとっては、その使い分けこそが困難であることが指摘されている（宇佐美 1995、2001)。そこで、ここでは、総発話文数に占める文末の丁寧体（「です」「ます」）の割合を「丁寧体率」と呼び、話者同士の関係、場面の違う会話における母語話者と非母語話者の丁寧体率を比較する。まず、「です」、「ます」それぞれの頻度と総発話文数に占める割合を以下の表６に示す。また、「です」「ます」を合わせた「丁寧体」の頻度とそれが総発話文数に占める割合である「丁寧体率」を次頁の表７に示す。また、「丁寧体率」は、次頁の図２にも示した。

表６　話者同士の関係、場面の違う会話における母語話者と非母語話者の「です」、「ます」の頻度と割合の比較

グループ名	母語話者・初対面		母語話者・友人		非母語話者*1・初対面		非母語話者*1・友人	
丁寧体の頻度（割合）	です	4,497 (31.6%)	です	518 (4.1%)	です	917 (16.6%)	です	61 (3.1%)
	ます	958 (6.7%)	ます	136 (1.1%)	ます	511 (9.2%)	ます	21 (1.1%)
	その他	8,366 (61.6%)	その他	12,121 (94.9%)	その他	4,108 (74.2%)	その他	1,894 (95.9%)
総発話文数	14,221(100%)		12,775(100%)		5,536(100%)		1,976(100%)	

*1　接触場面における非母語話者の発話のみを対象としている。

*2　丁寧体数の欄の括弧内は、発話文数に対する割合である。

表６を見ると、友人との会話における「です」「ます」の総発話文数に占める割合は、母語話者と非母語話者でほとんど差がなく、ともに５％以下と低いことがわかる。（χ^2検定

の結果、5%水準で有意差なし)。一方、初対面会話を見ると、母語話者の「です」が総発話文数に占める割合は 31.6%と高く、非母語話者の約 2 倍にのぼる。(χ^2検定の結果、母語話者と非母語話者の間に 1%水準で有意差が見られた。)逆に、非母語話者は、「ます」の使用率が 9.2%と母語話者よりも高くなっている。(χ^2検定の結果、1%水準で有意差が見られた。)すなわち、非母語話者の方が「ます」を相対的に多く用いていることが明らかになった。このことは、母語話者が、「行くんですか?」というところを、非母語話者は、「行きますか?」と言いがちであるというような報告等を支持しているように思われる。

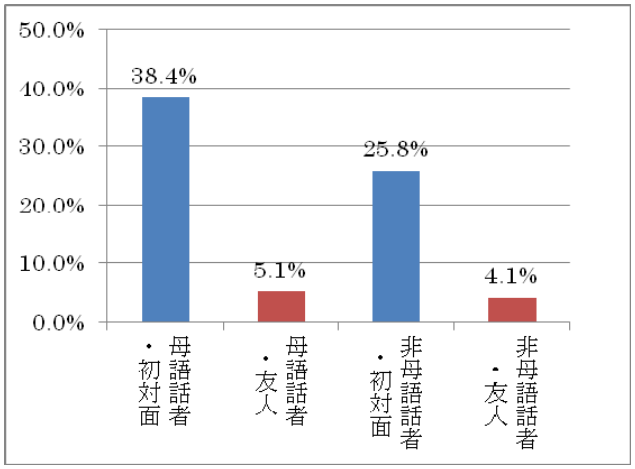
次に、「です」「ます」の頻度を合わせた「丁寧体率」について述べる。表 7 を見るとわかるように、母語話者も非母語話者も、友人との会話より、初対面会話で丁寧体を多く使っていることがわかる。友人同士の会話の丁寧体率は、母語話者、非母語話者ともに、約 5%と低いことでほぼ同様の傾向を見せる。友人同士の会話においては、χ^2検定を行った結果、母語話者と非母語話者の間に 5%水準で有意差は見られなかった。しかし、初対面会話を比較してみると、母語話者が約 40%の丁寧体率であるのに対して、非母語話者の丁寧体率は約 25%と低く、χ^2検定を行った結果、1%水準で有意差が見られた。

つまり、友人同士の会話においては、母語話者と非母語話者の丁寧体の使用に差はないが、初対面の会話においては、丁寧体の使用の違いが顕著であるということである。日本語において、初対面会話における「丁寧体」の適切な使用は、失礼のない円滑なコミュニケーションのために重要な要素の一つである。非母語話者の丁寧体率が、母語話者より有意に低いということは、語用論的に適切でない発話もありうる恐れもある。その点については、別途「定性的分析」と合わせて考察する必要がある。

表 7　話者同士の関係、場面の違う会話における母語話者と非母語話者の「丁寧体率」の比較

グループ名	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
丁寧体数	5,455	654	1,428	82
発話文数	14,221	12,775	5,536	1,976
丁寧体率	**38.4%**	**5.1%**	**25.8%**	**4.1%**

*接触場面における非母語話者の発話のみを対象としている。



5．まとめ

本稿では、『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 年版』の設計の趣旨と特性を紹介するとともに、本コーパスにおける動詞と副詞の高頻度語を

BCCWJ と比較した。また、話者同士の関係、場面の違う会話における母語話者と非母語話者の発話の「異なり語数」と「丁寧体率」の違いを明らかにした。「BTSJ 話し言葉コーパス」の会話は、諸条件を統制して収集し、相互作用研究に必須である発話の重なりや沈黙などが BTSJ のルールによってきめ細かく記述され、さらに、各会話参加者の社会的属性の情報がコーパス利用者の利便を考慮し、エクセルファイルにまとめられていることが特徴である。「総合的会話分析」（宇佐美 2008）という方法論では、「BTSJ 話し言葉コーパス」のこれらの特徴を活かして、本来は、ここに示した「定量的分析」の中身を「定性的分析」によってより詳細に分析・例示しながら、考察することをもって一研究と捉えることを主旨としている。定量的、定性的双方の分析を行って初めて「総合的会話分析」と言え、その目的である「人間の相互作用のメカニズムの解明」に貢献することができると考えるからである。ただ、今回は、「BTSJ 話し言葉コーパス」の設計と特性について概要を紹介するのが主旨であった。本コーパスを用いた本格的な語用論的、対人コミュニケーション的分析については、今後、稿を改めて発表していく。

謝　辞

本研究は、科学研究費補助金基盤研究（A)「自然会話リソースバンク構築による世界的教材共有ネットワーク実現のための総合的研究」(平成 23 年度～平成 26 年度、研究代表者：宇佐美まゆみ)による補助を得ている。記して感謝したい。

文　献

宇佐美まゆみ(1995)「談話レベルから見た敬語使用: スピーチレベルシフト生起の条件と機能」『学苑』662、pp.27-42. 昭和女子大学近代文化研究所

宇佐美まゆみ・嶺田明美（1995）「対話相手に応じた話題導入の仕方とその展開パターン: 初対面二者間の会話分析より」『名古屋学院大学日本語学・日本語教育論集』2、pp.130-145. 名古屋学院大学留学生別科 (日本研究プログラム).

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、pp.40-56、明治書院.

宇佐美まゆみ(2001)「『ディスコース・ポライトネス』という観点から見た敬語使用の機能 -敬語使用の新しい捉え方がポライトネスの談話理論に示唆すること-」『語学研究所論集』6、pp.1-29、東京外国語大学語学研究所

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」『講座社会言語科学 第 4 巻 教育・学習』、pp.150-181、ひつじ書房.

宇佐美まゆみ(2011)「基本的な文字化の原則(Basic Transcription System for Japanese: BTSJ)2011 年版」http://www.tufs.ac.jp/ts/personal/usamiken/btsj2011.pdf

林四郎(1975)「第二章　基本語彙はきめられるか」『新・日本語講座 1　現代日本語の単語と文字』、pp.37-54、汐文社.

関連 URL

宇佐美まゆみ研究室　http://www.tufs.ac.jp/ts/personal/usamiken/

宇佐美まゆみ監修(2011)『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』について http://www.tufs.ac.jp/ts/personal/usamiken/btsj_corpus_explanation.htm

宇佐美まゆみ（2012）「BTSJ 活用方法講習会の趣旨」
http://www.tufs.ac.jp/ts/personal/usamiken/btsj_koushuu_0_shushi.pdf

付表　『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』に収録されている会話グループとその概要

会話グループ番号と会話グループ名		会話の通し番号	データの特徴	データ数	総分数	音声付き
1	親しい同性友人同士(男女)の雑談	1-19	同性の友人同士の会話	19 会話	444 分 24 秒	
2	初対面と友人同士の女性の雑談	20-42	女性の、親しい友人同士と初対面の会話	23 会話	482 分 5 秒	
3	論文指導	43-52	教師と学生の面談の会話	10 会話	311 分	
4	女性同士の断りの電話会話	53-91	ある学生(女性)をベースに、電話で、先輩・同輩・後輩に依頼の電話をかけた会話	39 会話	78 分 35 秒	○
5	同性同士男女の依頼を含む電話会話	92-111	同性の友人同士の会話	20 会話	53 分 02 秒	
6	友人同士の女性の雑談	112-116	女性の友人同士の会話	5 会話	91 分 55 秒	
7	OPI インタビュー	117-120	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データ	4 会話	40 分	
8	韓国人学習者(中級)と日本人の初対面雑談	121-129	韓国人日本語学習者の接触場面データ	9 会話	249 分	
9	台湾人学習者(上級)と日本人の初対面雑談	130-141	台湾人日本語学習者の接触場面データ	12 会話	234 分 20 秒	
10	台湾人学習者(上級)と日本人の友人の雑談	142-151	台湾人日本語学習者の接触場面データ	10 会話	173 分 51 秒	○
11	初対面女性ベース雑談(接触、母語)その 1	152-160	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 32 秒	○
12	初対面女性ベース雑談(接触、母語)その 2	161-172	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	
13	初対面男性ベース雑談(性差、年齢差)	173-190	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	295 分 39 秒	○
14	初対面同性同士雑談(男、女)	191-206	20 代前半大学生・大学院生、初対面の雑談	16 会話	272 分 18 秒	○
15	友人同士女性雑談	207-209	20 代女性学生、親しい友人同士の雑談	3 会話	63 分 37 秒	

16	友人同士男女(雑談、討論)	210-233	10 代後半～20 代大学生友人同士の会話、ベース話者(男女同数)が、同性/異性の友人との雑談/討論という 4 通りの会話を行っている。	24 会話	401 分 16 秒	
17	友人同士男女間討論	234-238	20 代-30 代学生、友人同士の討論	5 会話	88 分 16 秒	
18	初対面女性討論	239-242	20 代女性、大学生・大学院生、初対面の討論	4 会話	44 分 33 秒	
19	友人同士女性誘い	243-250	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	
20	初対面女性雑談(母語・接触)	251-262	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	○
21	謝罪の会話	263-294	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	78 分 52 秒	○
計				294 会話	4041 分 39 秒 (約 67 時間)	

データ提供者は、下記の通り。(50 音順)。
李恩美、伊集院郁子、宇佐美まゆみ、カチマレク・ミロスワバ、北見奈津子、木林理恵、金銀美、木山幸子、黄瓊芸、施信余、鄭賢児、関崎博紀、蘇玉萍、高森絵美、張鈞竹、鄭榮美、藤田朋世、松本剛次、松本紫帆、宮武かおり、林君玲