

『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」に基づくものは多くは言えない。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」も増えてはいるが、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もあるが、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。その他、話し言葉コーパスに基づく分析は、講演などのストリート・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。「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しないが、「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、定量的分析に適した「文字化システム」に基づくものはほとんどない。すなわち、人間の相互作用の分析を企図し、会話の定量的分析に加えて、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。しかし、自然会話など、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、その後のコーディング(タグ付け)などにも膨大な時間と労力を要するため、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の共有していくことが不可欠である。また、会話データを共有していくためには、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き、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、定量的分析ができて文字化したデータを提供、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。

† usamime@uifs.ac.jp † nakamata@kyokyo-u.ac.jp

表3 超大規模コーパスプロジェクト：工程表

年	四半期	準備	収集	標準化	利活用	保守
2012	1Q	→ 計画立案	→ クローラ構築	→ 正規化技術調査	→ (後装) フォンエム	→ 注釈書作成
	2Q	→ 機材調達(初回)	→ クローラ運用テスト	→ 正規化技術調査	→ 音声列変換技術調査	→ 注釈書修正
2013	1Q	→ 機材調達(2回目)	→ クローラ本運用開始	→ 正規化技術調査	→ 音声列変換技術調査開始	→ 注釈書修正(2回目)
	2Q	→ 機材調達(3回目)	→ 運用規則見直し(初回)	→ 正規化技術調査	→ 音声列変換技術調査開始	→ 注釈書修正(3回目)
2014	1Q	→ 機材調達(4回目)	→ 運用規則見直し(2回目)	→ 正規化技術調査	→ 音声列変換技術調査開始	→ 注釈書修正(4回目)
	2Q	→ 機材調達(5回目)	→ 運用規則見直し(3回目)	→ 正規化技術調査	→ 音声列変換技術調査開始	→ 注釈書修正(5回目)
2015	1Q	→ 機材調達(6回目)	→ 運用規則見直し(4回目)	→ 正規化技術調査	→ 音声列変換技術調査開始	→ 注釈書修正(6回目)
	2Q	→ 機材調達(7回目)	→ 運用規則見直し(5回目)	→ 正規化技術調査	→ 音声列変換技術調査開始	→ 注釈書修正(7回目)
2016	1Q	→ 機材調達(8回目)	→ 運用規則見直し(6回目)	→ 正規化技術調査	→ 音声列変換技術調査開始	→ 注釈書修正(8回目)
	2Q	→ 機材調達(9回目)	→ 運用規則見直し(7回目)	→ 正規化技術調査	→ 音声列変換技術調査開始	→ 注釈書修正(9回目)

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話し言葉コーパスの設計の趣旨

「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、語用論的分析のために不可欠な情報のコーディングが行いやすい「基本的な文字化の原則」である BTSJ (Basic Transcription System for Japanese) によるトランスクリプトを用いる。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データベース（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¹⁾は 294¹⁾の相互作用的会話からなる。会話の総時間は 67 時間 20 分 9 秒、総語数は、781,022 語 (Mecab+UniDic、句読点等を除く実質的発話部) である。すべての会話には、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」がついており、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」も収録している。BTSJ トランスクリプトは、多くの人が活用しやすさを考え、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」の右側に分析項目の「コーディング」の列を追加し、コーディングすることによって、エクセルの機能を使って、話者ごとにソートして、話者の特徴を概観したり、特定の語彙の検索や、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。2007 年当初のエクセルをベースにしたシステムに、専用のマクロ機能を追加搭載して、2009 年に「BTSJ 入力支援・自動集計システムセット」を開発し、入力支援（対となる記号の自動入力、ショートカット機能付与等）や、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表示できるようにした。2011 年には、それに加えて、同じルールでコーディングした複数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」²⁾(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（今後「テキストファイル」も公開予定）。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録され、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⁾ 今回の分析対象は、修正中の 1 会話を除いた 293 会話のものである。

²⁾ 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、以下のように入力される。基本的に、ひとりの話者による「文」を成していること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「文」を成していること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「*」となっており、その後に「/」を記す。また、「発話内容」の列における「。」も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、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。

ライン番号	発話文番号	発話文終了	話者	発話内容
1	1	*	TF01	###しやべって下さい。
2	2	*	JF01	えっ、何をしやべるんですか？(笑い)。
3	3-1	/	TF01	沈黙(沈黙)ね、しやべらないよ。
4	4	*	JF01	<笑いながら>しやべらないよ。
5	3-2	*	TF01	おた私ずっとしやべってる感じだ(笑い)と思っちゃった。
6	5	*	TF01	あっ[[。
7	6	*	JF01]]あっ、お菓子食べる(仰)に...>K。
8	7	*	TF01	<そっか>。
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」「さえざり」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)を参照して頂きたい。

2. 4 『BTSJ 文字入力支援・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、本システムは、BTSJ による「文

字化」にかかる時間と労力を軽減する「文字化入力支援機能」と、BTSJで記されたトランスク립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(.xlb)」、「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。)本システムセットは、「BTSJ 活用方法講習会」の受講者には、無償で配布を行っている。

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。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	293 会話
総語数	783,103 語
異なり語数	12,076 語
TTR (異なり語数/総語数)	1.542%
Guiraud 値 (異なり語数/√総語数)	13.646
発話文数	90,593 文
1 文あたりの語数 (総語数/発話文数)	8.644 語
総時間	240,851 秒 (66 時間 54 分 11 秒)
1 文あたりの時間数 (総時間/発話文数)	2.659 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

「基本語集」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。基本語集の選定は、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

³ 解析は、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 特有の記号を除いた後、菜まめ(UniDic+mecab)を用いる形で形態素解析を行った。

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「文句なしの基本語集」とした 545 語であろう。その中に動詞は 78 語ある。BTSJ 話し言葉コーパスの高頻度語 20 の中には、「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」「書く」「取る」の 4 語も、BCCWJ でも、40 位以内に入っている。よって、これら 4 語は、書き言葉においても基本語に相当するとも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、「基本語集」(林 1975) との共通性も高いことが明らかになった。

一方、BCCWJ では 12 位に入っている「知る」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 0.78) 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	142.71	いる	109.73
2	する	119.93	する	60.51
3	ある	65.55	なる	48.69
4	行く	50.10	ある	47.20
5	思う	48.09	言う	30.43
6	やる	36.04	来る	22.83
7	いる	36.00	思う	20.35
8	なる	32.54	できる	13.29
9	来る	31.31	見る	18.14
10	分かる	23.70	行く	15.36
11	見る	23.22	しまう	9.69
12	違う	13.50	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	9.63	出る	8.01
17	書く	9.49	やる	7.80
18	知る	8.82	行う	5.95
19	考える	7.39	知る	5.79
20	取る	7.32	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 Nijjal-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 における高頻度の副詞上位 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つていて、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。すなわち、副詞の高頻度語は、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。

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」「もっと」「いっばい」「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」も、話し言葉と書き言葉では、かなり異なることがわかる。これらの結果から、副詞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よう。

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	167.78	そう	7.62
2	もう	36.32	どう	6.91
3	ちよつと	31.26	もう	5.14
4	こう	27.80	さらに	3.77
5	どう	25.28	やはり	3.24
6	やはり	25.08	まだ	3.04
7	まあ	23.06	よく	2.87
8	結構	16.17	少し	2.87
9	あまり	12.86	すぐ	2.52
10	多分	12.20	まず	2.52
11	全然	11.85	特に	2.48
12	まだ	8.50	まったく	2.37
13	よく	6.73	ちよつと	2.21
14	ずつと	5.08	すでに	2.14
15	色々	5.03	こう	2.08
16	なるほど	4.85	実際	2.02
17	うんうん	4.84	ほとんど	1.85
18	例えば	4.78	最も	1.74
19	一番	4.20	初めて	1.73
20	ちゃんと	3.81	もちろん	1.68

ただ、表3の上位の語に注目すると、「そう」が、双方のコーパスで1位であるという点と、「やはり」がBTSJ話し言葉コーパスで6位、BCCWJで5位という点で、両コーパスに共通しているように見える。しかし順位は同じでも、1万語あたりの頻度は、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」のように具体的な指示内容を持つ用法が多いことがわかって、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」(60%)、「やっば」(36%)であるのに対して、BCCWJでは、「やはり」(68%)、「やっぱり」(28%)、「やっば」(3%)となり、話し言葉と書き言葉の違いが顕著に現れてくる。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、単なる頻度の比較だけでなく、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だろう。

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』を用いた話用論的分析

ここでは、BTSJ話し言葉コーパス全体の語彙的特性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のが最大の特徴であり、特定の場面や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が重視されている。これは、多量のジャンルや属性のデータをあまり考慮せずに分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重視されていない点である。本節では、話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/非母語場面)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、それぞれの状況におけ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/非母語場面)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抽出し、「母語場面・初対面」「母語場面・友人」「非母語場面・初対面」「非母語場面・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「非母語場面・初対面」と「非母語場面・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・		非母語話者 ^{*1} ・	
	初対面	友人	初対面	友人
母語場面/ 非母語場面	22	18	24	10
母語場面/ 非母語場面	7	3	4	0
母語場面/ 非母語場面	3	3	0	0
母語話者・	64	54	28	10
非母語話者 ^{*1} ・	52	54	28	10
年齢	20代	20代 ^{*2}	20代	20代
	30代	30代	30代	30代
非母語話者の 出身			台湾	台湾
			中国大陸	中国大陸
非母語話者の 日本語レベル			超級	超級
			上級	上級
			中級	中級

*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代後半の数を含む。
*3 うち4名は中国大陸出身である。
*4 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	111,140	103,534	39,386	15,801
異なり語数	3,999	4,391	2,113	1,572
Guiraud 値	11.98	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。このことは、初対面会話という場面では、用いられる語彙に限られたものになりがちである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、丁寧体と普通体などのように使い分けられるからということは、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」を次ページの表7に示す。また、「丁寧体率」は、次ページの図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 4,254 (31.4%)	ます 934 (6.9%)	です 518 (4.1%)	ます 136 (1.1%)	です 917 (16.6%)	ます 511 (9.2%)	です 61 (3.1%)
その他	8,366 (61.7%)	その他 12,121 (94.9%)	その他 4,108 (74.2%)	その他 4,108 (74.2%)	その他 4,108 (74.2%)	その他 4,108 (74.2%)	その他 1,894 (95.9%)	その他 1,894 (95.9%)
総発話文数	13,554(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 丁寧体率の欄の括弧内は、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。(x²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31.4%と高く、非母語話者の約2倍にのぼる。(x²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。(x²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「行くんですか?」というところを、非母語話者は、「行きませんか?」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。次ページの表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%と低いことではほぼ同様の傾向を見せる。友人同士の会話においては、x²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、x²検定を行った結果、1%水準で有意差が見られた。

つまり、丁寧体の使用については、友人同士の会話ではなく、初対面の会話において、母語話者と非母語話者の違いが顕著であるということである。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にとって重要である。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得る恐れがある。その点については、今後、「定性的分析」と合わせて考察したい。

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	5,188	654	1,428	82
発話文数	13,554	12,775	5,536	1,976
丁寧体率	38.3%	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。

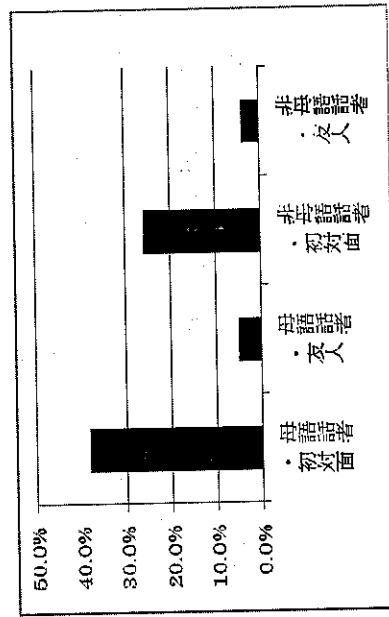


図2 各会話グループの丁寧体率

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、場面の違い会話における母語話者と非母語話者の発話における「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

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	総分數	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	53 分 28 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	167 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	○
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	271 分	○
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	

ーパス」の会話は諸条件が統制されて収録され、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収録されていることが特徴である。「総合的会話分析」(宇佐美 2008)という方法は、BTSJ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られるからである。ただ、それらの語用論的、対人コミュニケーション論的分析については、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(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)「相互作用と学習一ディスコース・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

百億語のコーパスを用いた日本語の語彙・文法情報 のブローファイリング

スルダノヴィッチ・イレナ (国立国語研究所・リュブリャナ大学)[†]
 スホメル・ヴィクト (マサリク大学言語処理センター)
 小木曾智信 (国立国語研究所)
 キルガリフ・アダム (レクシカルコンピューティング・リーズ大学)

Japanese Language Lexical and Grammatical Profiling Using the Web Corpus JpTenTen

Irena Srdanović (National Institute for Japanese Language and Linguistics/University of Ljubljana),
 Vit Suchomel (Natural Language Processing Centre, Masaryk University)
 Toshinobu Ogiso (National Institute for Japanese Language and Linguistics)
 Adam Kilgarriff (Lexical Computing Ltd./Leeds University)

1. はじめに

近年、一億語を超えた大規模な現代日本語書き言葉均衡コーパスが完成し、その大きなプロジェクトの成果として新しいアナノテーションツール、電子化辞書、コーパス検索ツールなどの日本語以外の様々な分野に応用できるようになる。今までのデータでは十分把握できず、抽出で必要な量的情報を増やす必要性が明らかになり、今までのデータでは十分把握できず、抽出できなかった言語的情報を得るために超大規模なウェブコーパス構築が始まった。こうした中、様々な言語でウェブコーパス作成の重要性が認識されてきて、多言語のためのTenTenと呼ばれるウェブコーパス群の構築が行われている。本論文において、まず新たに作成されたJpTenTenという日本語の100億語の超大規模なウェブコーパスを紹介する。このコーパスは、SpiderLing (Pomikalek and Suchomel 2012) などのツールでデータをクロールし、クリーニングを行った上で、MeCabとUniDic2 (小木曾ら 2011) で形態素解析し、短単位と長単位アナノテーションを付与した。コーパスはSketch EngineというレクシカルブラウザJpWaCを基にした語彙・文法ブローファイリングに搭載した。このツールは既に4億語の日本語コーパスJpWaCを基にした語彙・文法ブローファイリングを可能にしているが (Srdanović ら 2008)、本研究によって新たに可能になった成果は以下の通りである。

- 超大規模なコーパスを構築し、スケッチエンジンツールに載せた。その結果、今までできなかった言葉の組み合わせなどの言語情報を取り出せるようになった。
- 長単位と短単位のアノテーションを利用したこと、以前より統一された短単位のデータと、以前には存在しなかった長単位のデータが利用可能になった。
- 品詞タグだけでなく、UniDicの活用形および活用型等の英訳アナノーションを利用し、以前にはなかった活用形に関する詳細な情報を取り出せるようになった。
- 「文法関係ファイル」のデータを更に整備し、今まで取り出せなかった語と語の組み合わせおよびその振る舞いの情報が抽出できるようになった。

以上の外に、2語以上の共起抽出などの新しく開発した機能により、以前にはできなかった情報習得および表示ができるようになってきた。

本論文では、第2章においてコーパスの構築を紹介した上で、第3章においてコーパスのアノテーションおよび短単位と長単位の語彙ブローファイリングのメリットについて述べる。第4章は、新しい「文法関係ファイル」によって抽出できるようになった語彙・文法情報を紹介し、第5章では、具体的な例を取り出し、百億語の日本語のコーパスからどのような言語的情報が得られるかについて述べる。

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 の会話をしている。	24 会話	401 分
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	87 分
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
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	175 分
20	初対面女性雑談(母 語・接触)	251-262	日本語母話者同士の会話 と、日本語母話者と日本語 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	31 分32 秒
計				294 会話	4000 分 31 秒 (約 66 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、鄭賢兒、関崎博紀、蘇玉華、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、林君吟

[†] irena.srdanovic@ff.uni-lj.si