国立国語研究所学術情報リポジトリ

『BTSJ日本語自然会話コーパス2018年版』における一人称・二人称代名詞の使用実態

メタデータ	言語: Japanese
	出版者:
	公開日: 2022-08-03
	キーワード (Ja):
	キーワード (En):
	作成者: 宇佐美, まゆみ, 山崎, 誠
	メールアドレス:
	所属:
URL	https://repository.ninjal.ac.jp/records/3593

『BTSJ日本語自然会話コーパス2018年版』における一人称・二人称代名詞の使用実態

宇佐美まゆみ（国立国語研究所）、山崎 誠（国立国語研究所）

1．はじめに

国立国語研究所『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』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、年齢や性を条件統制して収集した会話データをまとめたもので、1会話20分程度の会話、333会話（約79時間）が収録されている。その主な基礎統計情報は、総発話文数104,489文、総語数922,444語、異なり語数13,469語、話者数（延べ666人、異なり435人）である。本稿では、この『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』を紹介するとともに、本コーパスを利用した研究の一例として、一人称・二人称代名詞の使用実態をまとめる。

2．『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』の構築の目的と特徴

近年、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、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「自然会話のコーパス」は、未だほとんどないのが現状である。しかし、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からこそ、自然会話分析研究を効率的に進めていくためには、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。このような現状を受けて、本コーパスは、自然会話データを研究者間で共有化することによって、自然会話データを用いた研究を促進することを企図している。

本コーパス構築のもう一つの趣旨は、未だ存在しない「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、以下の3点を重視した。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を収録する。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の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。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その知見を計量的にも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、大量データの形態素解析などのような言語形式の機械的な分析だけではなく、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（文脈）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（トランスクリプト・音声）2018 年版』の基本情報

このような趣旨のもと、宇佐美研究室[1]では、2007 年から、自然会話データを収集し、『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、『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きた[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、大幅に整備したものが『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、トランスクリプト中の記号などの表記を「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。以下に、本コーパスの基本情報をまとめる。

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』の語数等の基本情報

会話数	333 会話
延べ語数(Token)	920,169 語
異なり語数(Type)	13,331 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）	435 人

4．本コーパスにおける一人称・二人称代名詞の使用実態の計量的分析[3]

ここでは、本コーパスの中の母語場面と接触場面の初対面会話（男・女）、友人同士の会話（男・

[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/
[3] 紙幅の都合で、すべての集計結果を図表の形では示せないが、必要な結果は、「結果」や「考察」で言及した。

女)、合計228会話を対象として、一人称代名詞5種（わたし、あたし、うち、ぼく、おれ）と、二人称代名詞3種（あなた、きみ、おまえ）の各サブ・グループごとの総発話文数に占める頻度と割合、各サブ・グループ内での一人称・二人称代名詞の頻度と割合を算出し、その使用実態を分析した。以下の表2に、各条件の会話数・話者数・発話文数をまとめる。

表2　各条件の会話数・話者数・発話文数

		①母語場面・母語話者	②接触場面・母語話者	②接触場面・非母語話者
初対面	会話数	62	43	43
	話者数(全体)	124	43	43
	女性 相手の性別(女/男)	81 (72/9)	34 (34/0)	34 (34/0)
	男性 相手の性別(女/男)	43 (9/34)	9 (0/9)	9 (0/9)
	発話文数 (女/男)	22176 (15743/6433)	9789 (7369/220)	9221 (7001/2220)
友人	会話数	103	20	20
	話者数(全体)	206	20	20
	女性 相手の性別(女/男)	133 (104/29)	20 (20/0)	20 (20/0)
	男性 相手の性別(女/男)	73 (29/44)	0 (0/0)	0 (0/0)
	発話文数 (男/女)	45538 (31012/14526)	4375 (4375/0)	3986 (3986/0)

注：(1)対面会話、自然会話のみで集計(電話会話、ロールプレイを除く)。
(2)接触場面の会話は、母語話者、非母語話者とも同じもの。
(3)話者数は延べ人数で、全体(女性/男性)の形式で示す。

5．結果

ここでは、人称代名詞の使用実態を、量的、多角的に分析した結果を示す。

5.1　初対面と友人場面における人称代名詞の使用率の比較(全体)

以下の図1に、初対面と友人場面における各人称代名詞の使用率を示す。

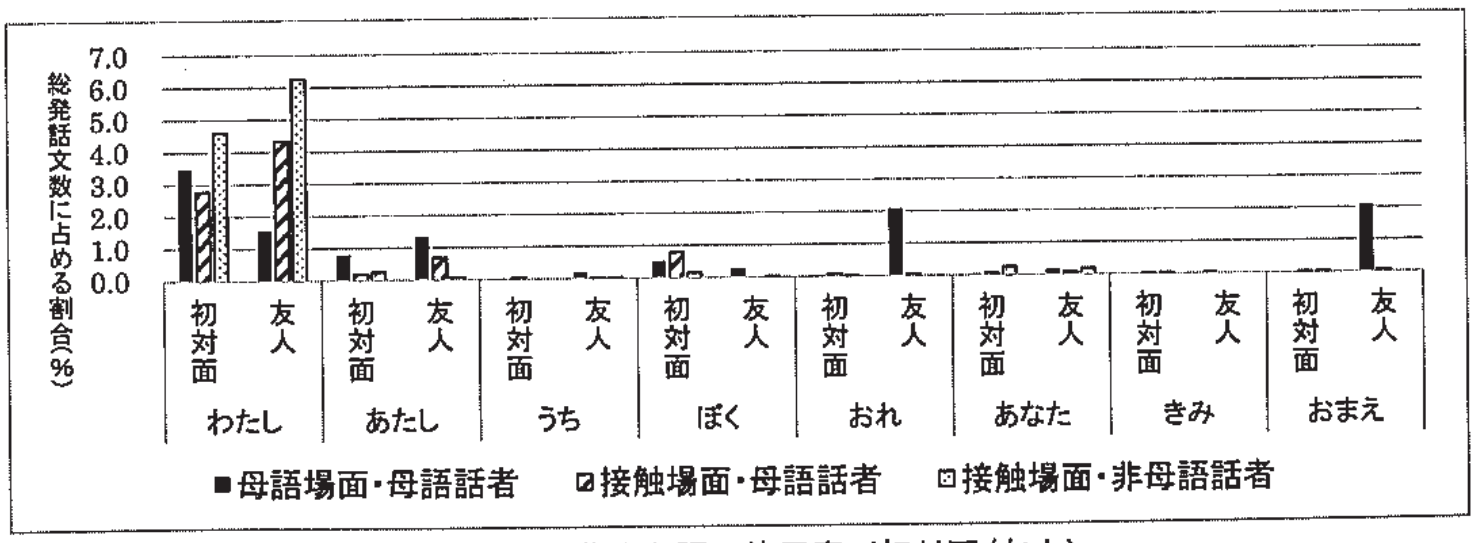

図1　人称代名詞の使用率（初対面/友人）

図１が示す顕著な傾向を以下に列挙する。

① 「わたし」の使用率は、母語場面の母語話者は、初対面場面のほうが友人場面より高い。

② 「わたし」の使用率は、接触場面においては、母語話者、非母語話者ともに、友人場面のほうが高い。また、非母語話者のほうが、母語話者より使用率が高い。

③ 母語話者は、「わたし」に代わり、初対面場面では「ぼく」（男性のみ）、「あたし」（引用部４回を除くと 女性のみ。友人場面より低い)を、友人場面では、「あたし」（女性のみ）「おれ」（男性のみ）を使用している。

④ 「あたし」は、女性が友人場面で用い、「ぼく」は、男性が初対面場面で相対的に多く用いるが、接触場面（対非母語話者）では対母語話者より多く用いている。

⑤ 二人称代名詞の使用率は、全般的に低いが、母語場面・母語話者の友人場面では、「おまえ」の使用率が高く、一人称の「おれ」と同様の比較的高い使用率（2%）を示している。

以下の図２に、同じデータを母語場面、接触場面における母語話者と非母語話者の言語行動を比較する観点から、「母語場面・母語話者」、「接触場面・母語話者」、「接触場面・非母語話者」の話者ごとの初対面、友人場面における人称代名詞の使用率を示した。

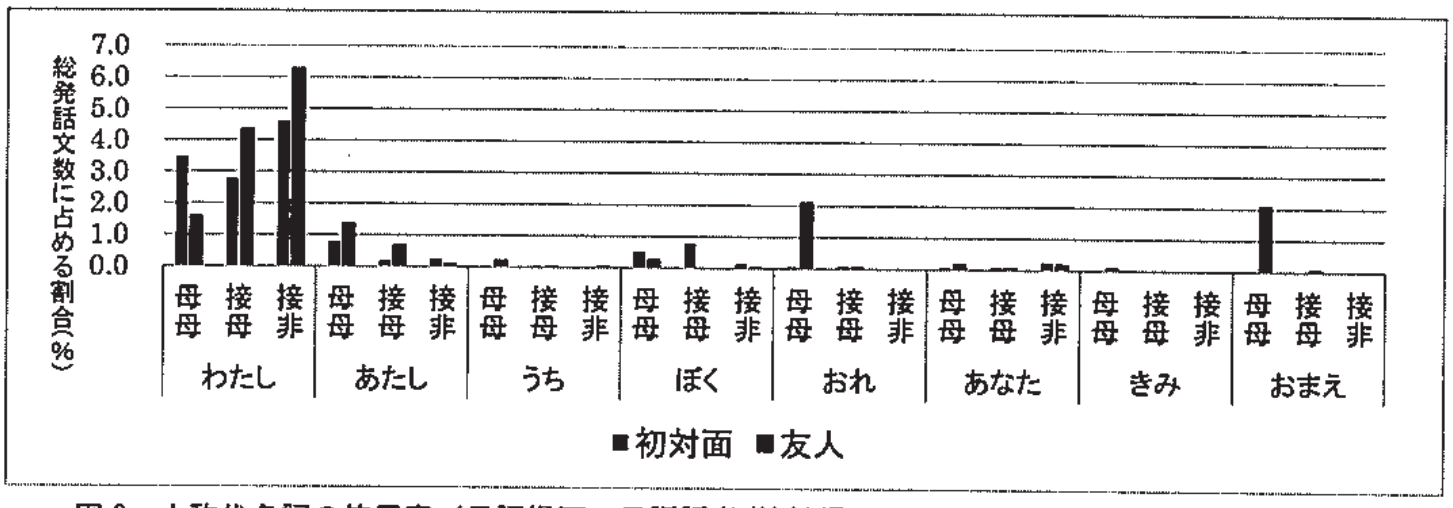

図２　人称代名詞の使用率（母語場面・母語話者/接触場面・母語話者/接触場面・非母語話者）

注　母母：母語場面・母語話者、接母：接触場面・母語話者、接非：接触場面・非母語話者（以下同様）

図２では、「わたし」については、母語場面の母語話者のみが他の 2 グループと異なる傾向を示していることがよりわかりやすく、また、上述の①~⑤の解釈を確認できる。

5.2　初対面と友人場面における人称代名詞の男女別使用率の比較（人称代名詞ごと）

以下の図 3-1～3-8 に、各人称代名詞の「男女別使用率」を人称代名詞ごとに示した。また、図４に、母語場面の母語話者のみ、各人称代名詞の男女別使用頻度を示した。

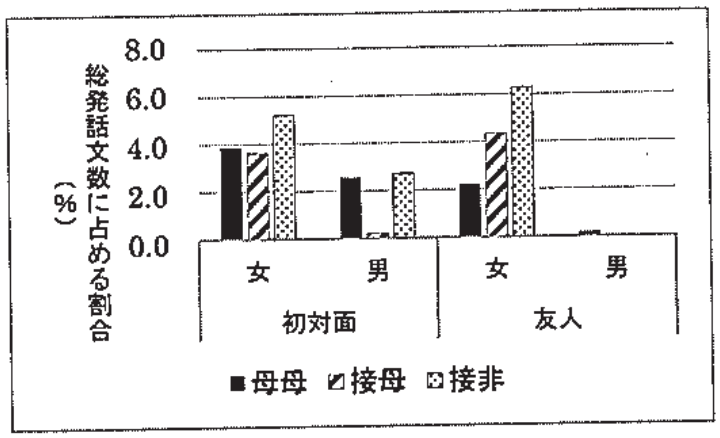

図 3-1　わたし

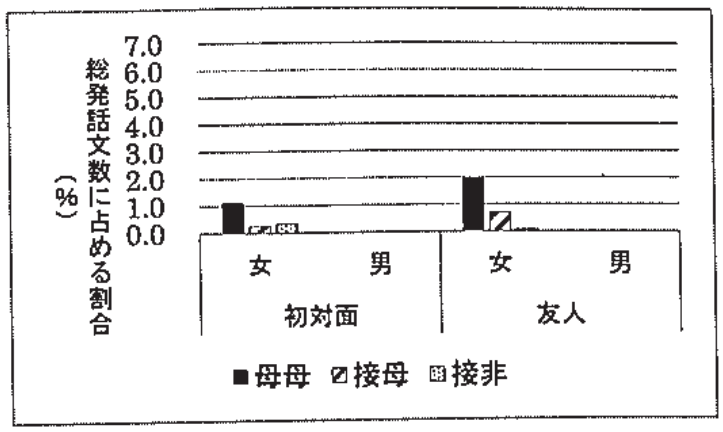

図 3-2　あたし

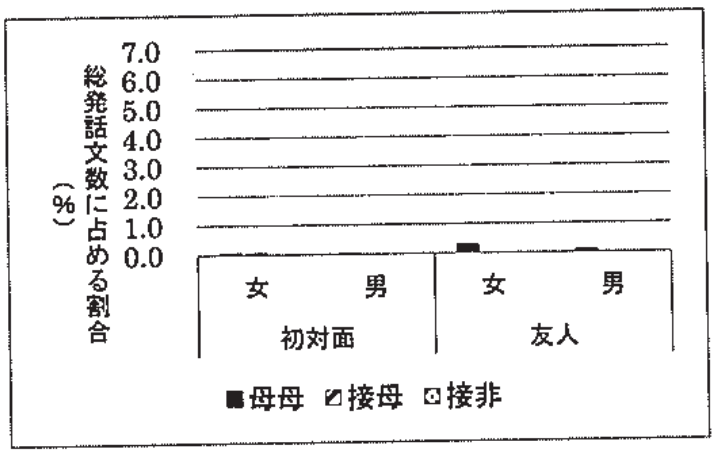

図 3-3　うち

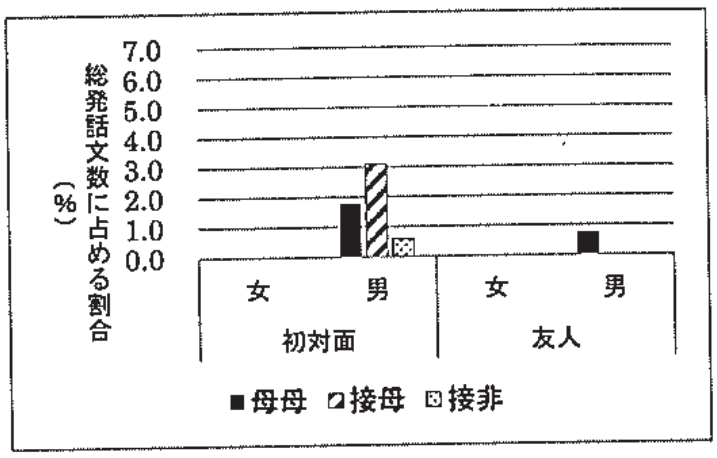

図 3-4　ぼく

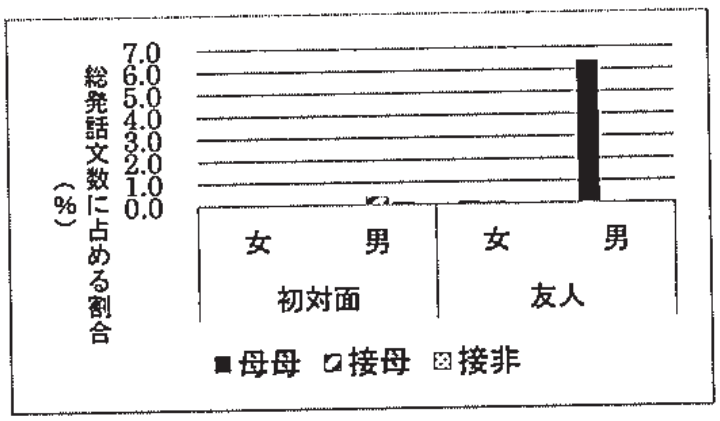

図 3-5　おれ

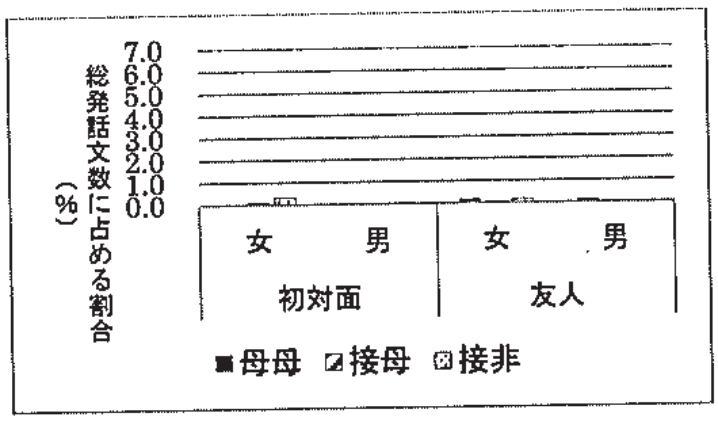

図 3-6　あなた

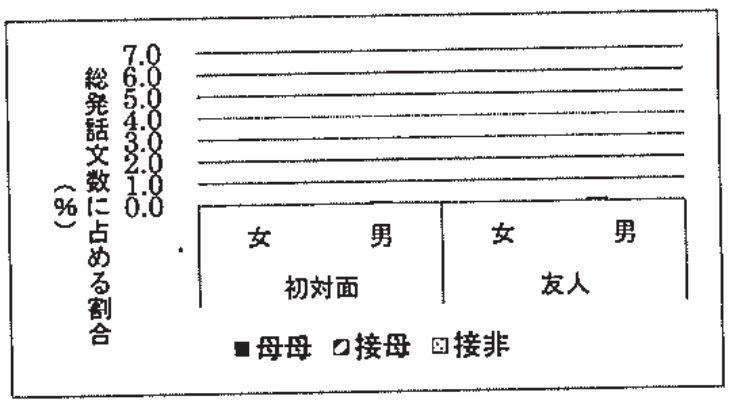

図 3-7　きみ

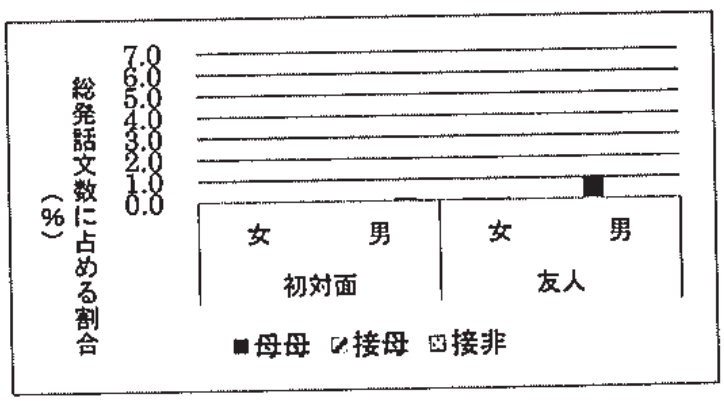

図 3-8　おまえ

図 3-1～3-8　各人称代名詞の使用率

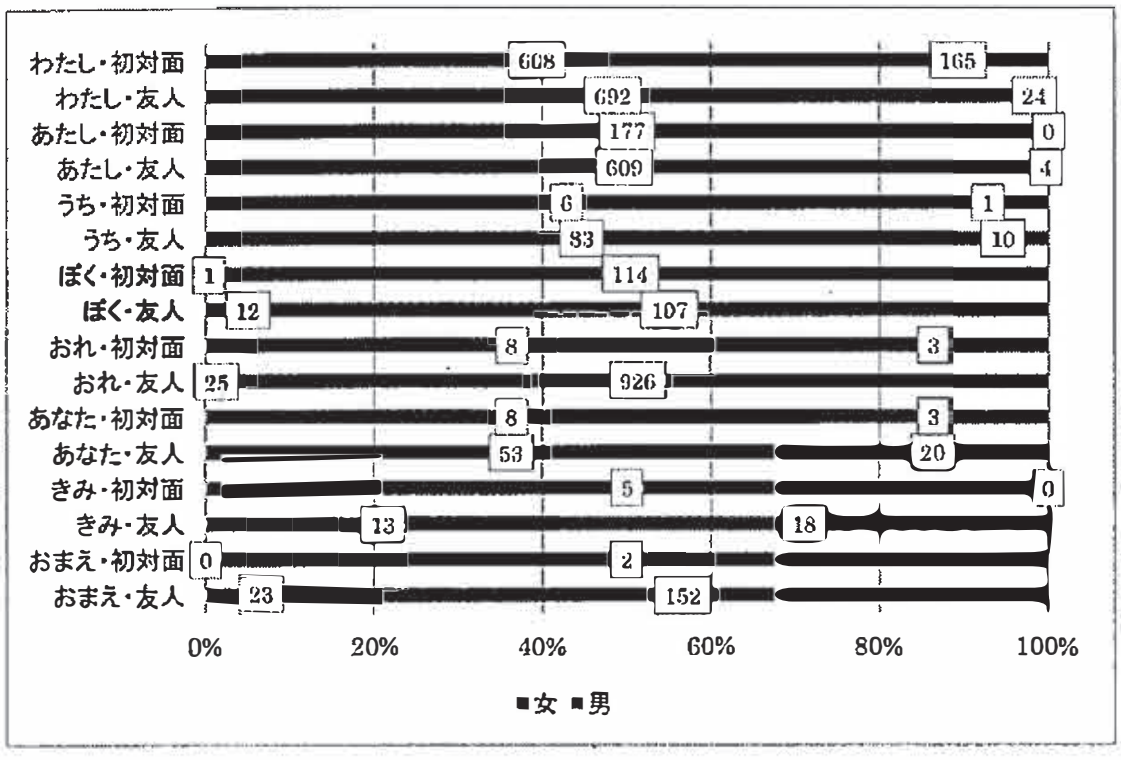

図 4　母語場面の母語話者における人称代名詞の男女別使用頻度

6．おわりに

以上、本コーパスの構築側として、まずは、333 会話の中から条件を統制して抽出した 228 会話という大量の会話における人称代名詞の全体的使用実態をまとめた。今後、個人の研究者が、本コーパスを利用する際には、本コーパス本体とは別に提供されている「　　」を活用して、条件を明確にして、各人の研究目的に適合する会話を抽出し、精選されたより少数の会話を用いて、語用論的、日本語学的観点から、質的分析を行うことによって、談話の流れや、発話のやりとりという「相互作用」を十分に考慮した研究を行うことなどが推奨される。すなわち、今後は、量的分析と質的分析双方からのアプローチによる研究成果に基づく「知見」を、分野として蓄積し、自然会話データという貴重なデータとそれを編んだ本コーパスのような自然会話コーパスの分析をより多角的に行うことによって、分析方法自体や解釈の質を高めていくことが必須である。本コーパスと、「総合的会話分析」の方法が、それらの一助となればと願っている。

【謝辞】

本研究は、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）、および JSPS 科研費 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）の成果の一部である。

【引用文献】

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）「相互作用と学習ーディスコース・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・山崎誠（2018）「『BTSJ 日本語自然会話コーパス（2018 年版）』構築の趣旨と特徴」『言語処理学会 第 24 回年次大会 発表論文集（2018 年 3 月）』: 420-423.