多様な様式を網羅した会話コーパスの共有化

伝 康晴 (千葉大学文学部/国立国語研究所言語資源研究系)[†]
土屋 智行 (国立国語研究所言語資源研究系)
小磯 花絵 (国立国語研究所理論・構造研究系)

Sharing of Conversation Corpora That Cover Diverse Styles and Settings

Yasuharu Den (Faculty of Letters, Chiba University/Dept. Corpus Studies, NINJAL)
Tomoyuki Tsuchiya (Dept. Corpus Studies, NINJAL)
Hanae Koiso (Dept. Linguistic Theory and Structure, NINJAL)

1. はじめに

近年、書き言葉コーパスの構築は飛躍的な発展を見せている。国立国語研究所では、1 億語を超える規模の『現代日本語書き言葉均衡コーパス』[*1]を開発し、さらに 100 億語を超える規模の Web コーパスの開発を目指している。これに対して、話し言葉コーパスは、音声収録・転記など開発初期段階での負担が大きく、学会講演や模擬講演などの独話を中心とする『日本語話し言葉コーパス』[*2]を除いて、大規模なものは存在しない。とくに日常の言語行動の中心である会話に関しては、個々の研究プロジェクトごとに小規模なデータを独自に収集・利用している状態を脱していない。

これに対する一つの解決策として、既存の会話コーパスの共有化という方式に着目する。小規模データを所有する研究機関は多くあり、それらは音声収録・転記の段階を終え、負担の大きい初期のハードルをクリアしている。これらのコーパスを共有すれば、研究に利用できる会話データの量は従来よりも飛躍的に増加する。しかし、これらのコーパスでは、転記方式は不統一であり、また、韻律情報や発話機能など会話研究に必要な基本情報は必ずしも完備していない。そこで、本研究では、これらの基本情報に関する共通のアノテーションを施し、相互利用可能な形で会話コーパスを共有する方法を考案する。

本稿では、上記の目標を達成するために立案したプロジェクトの概要、および、転記方式に関する予備的な調査結果について述べる。

2. プロジェクトの概要

1 節の目標を達成するため、国立国語研究所独創・発展型共同研究「多様な様式を網羅した会話コーパスの共有化」（リーダー：伝康晴・2011 年 11 月～2014 年 10 月）を開始した。本プロジェクトは、2.1 の研究組織のもと、2.2 の研究計画で遂行する。プロジェクトの進捗状況

[†] den@cogsci.l.chiba-u.ac.jp
[*1] http://www.ninjal.ac.jp/kotonoha/
[*2] http://www.ninjal.ac.jp/csj/

表 1　プロジェクト内で共有可能な会話コーパス（予定）。各コーパスの出典については付録 A を参照

名称	参与者数	関係性	様式	内容
千葉大 3 人会話	3 人	友人	対面	雑談
CSJ	2 人	初対面（一方はプロ）	対面	インタビュー
FGI	4 人	初対面 + プロ/アマチュア	対面	インタビュー
言語接触場面 3 人会話	3 人	知人（非母語話者 1 名）	対面	雑談
新聞販売店会話	2 人	店員と顧客	非対面	ビジネスコール
早稲田大自由対話	2 人	友人（ゼミ配属前/後）	対面/非対面	雑談
JPN	2〜3 人	友人/家族	対面	雑談
作業療法会話	主に 2 人	療法士とクライエント	対面	作業療法
宇都宮大音声対話	2 人	友人	非対面	課題指向
三重大地図課題対話	2 人	知人/初対面	対面	課題指向
タングラムパズル対話	2 人	知人/初対面	対面	課題指向
ロゴ積み木対話	2 人	知人/初対面	対面	課題指向
北大 2 人会話	2 人	先輩後輩（非母語話者 1 名）	対面	雑談

および研究成果は、「会話コーパス」プロジェクトホームページ[*3]から随時発信していく。プロジェクト終了時には以下のものを公開する予定である。

- 基本情報の仕様をまとめたマニュアル
- 共有コーパスに付与されたアノテーション[*4]

2.1　研究組織

研究組織は著者たち以外に 9 名の共同研究者・研究協力者からなり、その専門領域は会話分析・談話分析・日本語教育学・日本語学・認知心理学・音声言語情報処理など多岐にわたる（一覧はプロジェクトホームページを参照）。いずれも会話データをみずから収集し、研究に利用している研究者たちであり、本プロジェクト内でコーパスを共有し、共有化に伴う問題点を解決し、共有化の利点を検証するという趣旨に賛同していただいている。

2.2　研究計画

本プロジェクトでは、1 節の目標を達成するため以下のことを行なう。

2.2.1　会話コーパスの調査

会話の諸現象の普遍性と多様性をとらえるために、参与者数・関係性・様式・内容などがさまざまに異なるコーパス群を対象とする。そのためまず、各メンバーが所有するデータ（表 1）を対象として、転記方式や音質・付加情報などを調査し、共有化に伴う問題点を洗い出す。現在、転記方式に関する予備的な調査を進めており、その結果について 3 節で報告する。

2.2.2　共通の基本情報の仕様策定

2.2.1 の調査に基づき、共通に付与できる基本情報（表 2）の仕様を策定し、マニュアルとしてまとめる。

[*3] http://www.jdri.org/kaiwa/

[*4] 公開の可否および形態は各コーパスごとに異なる。

表 2　共通に付与する基本情報（予定）

基本情報	説明	予想される問題点
転記	発話内容の文字化	非流暢性のマークアップや基本単位の不統一
形態論情報	単語分割・品詞	自動解析の可否
韻律情報	発話末音調など	音質によっては主観的付与に限定される
発話機能	談話行為・宛て先など	従来は課題指向対話がおもな対象
局所構造	隣接ペア・発話交換構造	多人数会話への拡張
連鎖構造	先行連鎖・修復連鎖など	これまで明確に策定された基準はない

2.2.3　共通の基本情報の付与

2.2.2 で策定された仕様に基づき、各コーパスに基本情報を付与する。作成したデータはサーバ上で管理し、メンバー間で共有する。

2.2.4　共有コーパスの基礎的分析

2.2.3 の共有コーパスを用いて、各メンバーがこれまで行なってきた研究テーマ（話者交替・あいづち・連鎖構造・成員カテゴリー化など）に関する基礎的分析を行なう。とくに、様式が異なるコーパス間の普遍性と多様性を明らかにする。これによって、多様な様式の会話コーパスを共有することの利点を検証する。

2.2.5　プロジェクト外の会話コーパスの調査

本プロジェクトのメンバーが所有する以外の会話コーパスについて、2.2.1 と同様の調査を行ない、本手法で共有化できるデータがどれくらいあり、どの程度の多様性を網羅できるか調査する。これによって、より大規模な会話コーパスを設計する際の指針とする。

3.　転記方式に関する調査

研究計画の 2.2.1 の最初のステップとして、各メンバーが利用しているコーパスの転記テキストの断片（数分程度）を収集し、さまざまな観点から比較した。表 3 にその概略を示す。また、いくつかのコーパスにおける転記テキストの例を付録 B に示す。

以下、項目ごとに調査結果の概要を述べる。

3.1　転記単位

転記テキストをどのような単位で分割するかについて、CSJ のように明確に定めている（200 ミリ秒以上の無音もしくは明示的な文末表現で分割）場合もあるが、多くのコーパスでは転記単位は不明確であった。転記単位の調査は今後進めたい。

3.2　レイアウト

転記テキストのレイアウトは、多くの場合、1 行に（ないし数行にわたって）1 つの発話単位を記し、発話単位ごとに行を重ねる一般的な書式であったが、中には、参与者ごとに列を区切り、複数の話者が同時に産出した発話単位を同じ行に記すことでタイミングの同時性を表わせるように工夫しているものもあった（付録 B：言語接触場面 3 人会話）。

表 3　転記方式の比較

コーパス	時間情報	文字表記	転記基準
千葉大 3 人会話	単位開始/終了時間	漢字かな混じり	CSJ 方式（簡略版）
CSJ	単位開始/終了時間	基本形・発音形併記	CSJ 方式
FGI	単位開始/終了時間	漢字かな混じり	独自方式
言語接触場面 3 人会話	なし	漢字かな混じり	独自方式
新聞販売店会話	単位内・単位間休止	漢字かな混じり	会話分析方式
早稲田大自由対話	単位開始/終了時間	基本形・発音形併記	CSJ 方式（簡略版）
JPN	単位間休止	ローマ字	Santa Barbara 方式
作業療法会話	なし	漢字かな混じり	独自方式
宇都宮大音声対話	単位開始/終了時間・単位内休止	基本形・発音形併記	談話タグ WG 方式
三重大地図課題対話	単位開始/終了時間・単位内休止	ひらがな	千葉大地図課題方式
タングラムパズル対話	単位開始/終了時間	ひらがな	独自方式
ロゴ積み木対話	単位内・単位間休止	漢字かな混じり	会話分析方式
北大 2 人会話	単位内休止・単位間休止	漢字かな混じり	会話分析方式

コーパス	非言語音	非流暢性	音調	重複位置
千葉大 3 人会話	笑	フィラー・語断片・延伸	（別ファイル）	なし
CSJ	笑・咳・息	フィラー・語断片・延伸	（別ファイル）	なし
FGI	笑	なし	上昇	相づちのみ
言語接触場面 3 人会話	笑・咳	延伸	上昇	あり
新聞販売店会話	笑・咳・息	語中断・延伸	上昇・下降・継続	あり
早稲田大自由対話	笑	なし	なし	なし
JPN	笑	語断片・延伸	上昇・下降・継続	あり
作業療法会話	笑	延伸	上昇	なし
宇都宮大音声対話	笑・息	フィラー・語断片	なし	なし
三重大地図課題対話	笑	なし	上昇	あり
タングラムパズル対話	笑	延伸	なし	あり
ロゴ積み木対話	笑	語中断・延伸	上昇・下降・継続	あり
北大 2 人会話	笑・息	語中断・延伸	上昇	あり

3.3　時間情報

いくつかのコーパスでは、転記単位ごとに開始・終了時間が与えられていた（付録 B : CSJ）。また、開始・終了時間が与えられていない場合でも、発話単位内・単位間に生じた休止の長さを秒や記号（`(.)`：短い休止）で記しているものがあった（付録 B：新聞販売店会話）。

3.4　文字表記

漢字仮名混じりによる表記が半数を占めていたが、実際に発音された音列を仮名で併記しているもの（付録 B : CSJ）や、ローマ字や仮名のみで表記しているものもあった（付録 B : JPN）。

3.5　転記基準

公刊されている転記基準に準拠しているものと、独自の方式で転記しているものがあった。公刊されている転記基準としては、CSJ 方式 (小磯他 2006) に準拠したもの（付録 B : CSJ）と

会話分析方式 (Jefferson 2004) に準拠したもの（付録 B：新聞販売店会話）が多く見られた。これらはともに非言語音の転記や非流暢性に関する豊富なマークアップを規定しているが、その表現の仕方はまったく異なる。

3.6 非言語音の転記の有無

多くのコーパスでは、参与者の笑い声の転記が行なわれていた。咳や息については、省略しているものが多かった。

3.7 非流暢性のマークアップの有無

音の延伸（たとえば「夕刊のほう:::」）は多くのコーパスで特別な記号によりマークアップされていた。フィラー（「あの」「えーと」など）や語断片（語としての形をなしていない音列）は CSJ 方式に基づく一部のコーパスのみでマークアップされていた。会話分析方式では、語が途中で中断された場合のみ（「え-」など）マークアップされていた。

一般に、音の延伸以外はマークアップされていないコーパスが多かった。ただし、マークアップはなくとも、転記としては記されている場合がほとんどであった。

3.8 音調のマークアップの有無

多くのコーパスは上昇調を転記しており、上矢印（↑）（付録 B：言語接触場面 3 人会話）やクエスチョンマーク（?）（付録 B：新聞販売店会話、JPN）などの記号を用いていた。下降・継続の音調をマークアップしているものは、会話分析方式に準拠した一部のコーパスに限られた。ただし、他のコーパスの中にも、韻律情報アノテーションとして別ファイルで音調の情報を与えているものもあった。

どの範囲の音調を転記テキスト中にマークアップするのが利便性がよいか、調査する必要がある。

3.9 重複位置のマークアップの有無

会話分析方式を中心に、複数の話者の発話の重複をマークアップしているものが多く見られた。一方、CSJ 方式に準拠したコーパスでは重複位置はマークアップされていなかった。これらのコーパスでは、単語（千葉大 3 人会話）や音韻（CSJ）ごとに開始・終了時間が与えられており、それらの情報から自動的に重複位置をマークアップすることが原理的には可能である。

重複位置のマークアップ方法もさまざまで、重複箇所や範囲を記号のみで示す方法（付録 B：言語接触場面 3 人会話）と、記号によるマークアップとともに、字下げによって発話開始位置をそろえ視覚的によりわかりやすくする方法（付録 B：新聞販売店会話、JPN）があった。

3.10 まとめ

以上のように、転記方式はコーパスごとにさまざまであるが、CSJ 方式や会話分析方式を中心にいくつかのグループにまとめられそうである。今後、これらのグループ内での細部の差異の調査と、これらのグループを超えた共通の転記基準の策定が可能かなど、調査を進めたい。

4. おわりに

本稿では、既存の会話コーパスの共有化を目標とする研究プロジェクトの概要、および、転記方式に関する予備的な調査結果について述べた。本プロジェクトによって、既存の会話コーパスの共有化の可能性・有効性が示されれば、今後、より広い範囲でコーパス共有を試みたい。同時に、既存のコーパスでは網羅できない様式や内容に関しては新規のデータ収集も必要となろう。将来的には、これらのことを目標に、大規模会話コーパスの構築へとつなげたい。

謝辞　本研究は国立国語研究所独創・発展型共同研究「多様な様式を網羅した会話コーパスの共有化」（リーダー：伝康晴）による成果である。調査に協力していただいたメンバー各位に感謝する。

参考文献

Jefferson, Gail (2004). “Glossary of transcript symbols with an introduction.” Gene Lerner (Ed.), *Conversation analysis: Studies from the first generation*. Amsterdam: John Benjamins. pp. 13–31.

小磯花絵・西川賢哉・間淵洋子 (2006).「転記テキスト」　『国立国語研究所報告書 124: 日本語話し言葉コーパスの構築法』　pp. 23–132.

関連 URL

「会話コーパス」ホームページ：`http://www.jdri.org/kaiwa/`

付録 A. コーパスの出典

※公刊物がない場合は問い合わせ先を記載

千葉大 3 人会話

Den, Yasuharu, and Mika Enomoto (2007). "A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation." Toyoaki Nishida (Ed.), *Conversational informatics: An engineering approach*. Hoboken, NJ: John Wiley & Sons. pp. 307–330.

CSJ

前川喜久雄 (2004).「『日本語話し言葉コーパス』の概要」 日本語科学, 15, pp. 111–133.

FGI

Morimoto, Ikuyo, Kana Suzuki, Etsuo Mizukami, and Hiroko Otsuka (2008). "Categorization in Japanese group discussion: Its advantages and disadvantages." *Proceedings of the 15th World Congress of Applied Linguistics*. Essen, Germany.

言語接触場面 3 人会話

大場美和子 (2011).「内的場面と接触場面における三者自由会話への参加の調整—談話・情報・言語ホストの役割の分析—」 博士論文（未公刊）, 千葉大学大学院人文社会科学研究科.

新聞販売店会話

Suzuki, Kana (2010). "Other-initiated repair in Japanese: Accomplishing mutual understanding in conversation." Unpublished doctoral dissertation, Graduate School of Intercultural Studies, Kobe University.

早稲田大自由対話

菊池英明（早稲田大学人間科学学術院）

JPN

大野剛（カナダ・アルバータ大学)、鈴木亮子（慶應義塾大学経済学部）

作業療法会話

長岡千賀（京都大学こころの未来研究センター）

宇都宮大音声対話

Mori, Hiroki, Tomoyuki Satake, Makoto Nakamura, and Hideki Kasuya (2011). "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics." *Speech Communication*, 53, pp. 36–50.

三重大地図課題対話

吉田悦子 (2002).「日本語名称なし地図課題対話コーパスの概要と転記テキストの作成: 報告」 人文論叢（三重大学人文学部文化学科研究紀要）, 19, pp. 241–249.

タングラムパズル対話

吉田悦子（三重大学人文学部）

ロゴ積み木対話

谷村緑、吉田悦子、竹内和広 (2009).「課題遂行対話におけるグラウンディング成立の記述方法の検討—日本人英語学習者の場合」 社会言語科学会第 23 回大会論文集, pp. 162–165.

北大 2 人会話

山本真理（北海道大学大学院国際広報メディア・観光学院）

付録 B. 転記テキストの例

言語接触場面 3 人会話

行	NS1	NNS	NS2
17			あじゃ相当寒いーんです
18	/ん/		よね↑/2 月って一番
19		[[んー]]	寒いんですか [ねやっぱ]
20	2 月とか寒かった		
21		2 月は一寒かった	
22		ちょっと雪が降ってました	

CSJ

```
0147 00252.842-00255.979 R:
ほんの                                & ホンノ<H>
三十分も                              & サンジュップンモ
いなかったと                          & イナカッ(? タ)ト
思うんですけど                        & オモウンデスケド<H>
0148 00255.410-00256.211 L:
(F うーん)                            & (F <VN>)
0149 00256.272-00257.115 R:
乗り換えで                            & ノリカエデ
0150 0150 00257.922-00259.109 L:
やっぱり                              & ヤッパリ<H>
(F そのー)                            & (F ソノー)
```

新聞販売店会話

```
 8 C    え：：：と日経が↑間違って入＜ってます＞。
 9 A    あ：：：＞そうですk-＜<<↑夕刊のほう：：[：です：ね？＝
10 C                                          [はい。      ＝はい。
11 A    え：と、(0.3)  ↑朝日の↓ほう：：：
12      (.)
13 C    はい、そ [うです。
14 A             [↓゜はい゜、入れればえ-（.） [よろしいんですね？
15 C                                          [はい
16      ↓はい。＝
```

JPN

```
216 H: .. jibun dake <X un X>,
217    yasunde,
218    warui na,
219    tte no mo,
220    an ja nai [no]?
221 R:           [un].
222    ... [2 sore de  2],
223 H:     [2 soo ieba 2],
224    goorudenuiiku da <X mon X> na=.
```