

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

著者	吉田 悦子
雑誌名	人文論叢 : 三重大学人文学部文化学科研究紀要
巻	19
ページ	241-249
発行年	2002-03-25
その他のタイトル	ニホンゴ メイショウ ナシ チズ カダイ タイワ コーパス ノ ガイヨウ ト テンキ テキスト ノサクセイ ホウコク An Outline and Transcription Text of Japanese Unlabelled Map Task Dialogue Corpus : A Report
URL	http://hdl.handle.net/10076/1905

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*

吉田悦子

要旨 本稿は2001年三重大学において収録された「日本語名称なし地図課題対話コーパス」〔以下、「日本語コーパス」〕の作成について報告するものである。今回の実験の結果、4人の話者による8対話が2チャンネル音声データとして収録され、千葉大学で開発されたツールを利用して転記作業がおこなわれ、時間情報を付与した転記テキストを得ることができた。さらにコーパス利用のためのツールによって、転記テキストを参照しながら音声を再生することが可能となり、Windowsの環境でコーパスを利用した対話分析をおこなえる基盤ができたことは大きな成果であるといえる。

1 はじめに

本稿は2001年に収録された「日本語名称なし地図課題対話コーパス」〔以下「日本語コーパス」〕の概要および収録対話の転記テキストの特徴について報告するものである。

1.1 目的

日本語コーパス作成の目的は、1. エディンバラ大学で収録された英語名称なし地図課題対話コーパス〔以下、「エディンバラコーパス」〕と同じ規模、デザインになるよう考慮して、日英語対照研究を可能にする「より自発的で自然な」日本語の対話コーパスを作成し、共有できるデータとして電子化すること¹、2. すでに千葉大学で完成した日本語コーパス〔以下、「千葉コーパス」〕では出現が限られている指示表現を誘発する環境を設定し、その分布と用法を記述し、被験者間の相互作用的なコミュニケーション活動として機能していることを例証すること²、である。

1.2 予備実験における問題点とその解決方法

吉田（2001）では、予備実験および転記テキスト作成作業全体を通して、実質的かつ技術的な問題点として、音声情報の欠落および不正確さ、収録条件の不統一性、転記テキスト記述上の不正確さが指摘された。この3点の問題点をふまえて、より信頼できうる対話収録とコーパス作成のための解決策として以下のように提示する。

1. 2チャンネル音声データの確保
2. 必要最低限の厳密な音声情報の記述
3. エディンバラコーパスとの収録条件の統一
4. 転記テキストの基準の確立

また、日本語コーパス作成の背景についての詳細は吉田（2001）を参照されたい。

2 地図課題について

地図課題は英語、日本語共通のものを使用した³。地図課題の種類は4種類ある。今回の日本語コーパスは、エディンバラ、千葉大のオリジナルコーパスの基本設計を踏まえながらも、あえて地図上の目標物の名称を省いた地図課題を用いた。その結果、被験者がより自由に地図上の目標物を指示する余地が生まれてくることになり、指示の種類と分布形態に関してオリジナルコーパスとは明らかな違いが出てくることが予想される。千葉コーパスで使用した地図課題の概要は以下の通りである。（堀内他：1997：3）

地図課題は二人の実験参加者により共同で達成される課題である。まず、各参加者には一人一枚ずつ地図が与えられる。そのとき、一人の実験参加者（以後、情報提供者：Instruction Giver）の地図には経路が記されているが、もう一人の実験参加者（以後、情報追随者：Instruction Follower）の地図には経路が記されていない。二人の実験参加者に課される課題は、相手の地図が見えないように向かい合い、お互いに会話を交わしながら、情報提供者の地図上の経路を情報追随者の地図上に再現することである。なお、情報提供者の地図には経路以外に出発地点と目標地点、そしていくつかの目標物とその名称がえがかれている。一方、情報追随者の地図には、出発地点と目標物およびその名称だけが描かれており、目標地点と経路は描かれていない。二つの地図は完全に同一ではないが、そのことは実験の前にあらかじめ実験参加者に明示的に教えられている。しかし、二つの地図がどのように異なっているかは伝えられておらず、課題の遂行過程で解決すべき問題の一つとされる。（アンダーラインは著者）

したがって、予備実験同様、オリジナルコーパスで用いた地図課題と今回の実験で用いた地図課題との重要な違いは、「地図上の目標物に名称をつけていない」という点にある。（図1参照：地図番号--1）

3 収録の実施

2001年3月7日、津市島崎町の防音設備のあるスタジオ室内において話者4名、収録対話数8対話分の収録がおこなわれた。

3.1 収録設備

スタジオ内では、地図を置く机が向かい合って設置され、机と机の間には仕切りが設けられている。実験参加者は、この仕切りをはさんで向かい合って座る。仕切りは、お互いの顔のみが見える高さにあるが、相手の手元と地図は見えないように設置されている（視認可能）。しかし、身振り手振りを相手に示すことは禁じられている。

3.2 録音方法

スタジオにおける全ての録音作業は、録音技師（ホールオペレーター）が担当する。録音にはスタンド式のマイクを利用し、参加者の口元より10cmくらいの位置にマイクがくるように高さを調節する。マイクから収録された対話の音声は、それぞれを同時にDATでステレオ録

この箇所の図版は著作権の
関係で削除した

図 1：地図課題の例

音する（左チャンネル／Lに情報提供者／Giverの音声、右チャンネル／Rに情報追随者／Followerの音声を録音する）。その後、ノイズの除去と声の音量調節をおこない、DATからCD-Rに焼き付けた。

3.3 実験参加者

実験参加者は21歳から24歳の三重大大学の学生で、男女それぞれ2名ずつで4人組（A, B, C, Dとする）を構成した。A, Bは男性ペア、C, Dは女性ペアである。各ペア内の二人はお互いによく知っている間柄である（親近性がある）が、他方のペアのメンバーとはどちらの話者とも互いに面識がない（親近性がない）⁴。それぞれの実験参加者は、情報提供者と情報追随者、両方の役割で親近性のあるパートナー、および親近性のないパートナーとそれぞれ一度ずつ4回対話をおこなう。用いる地図は、情報提供者の場合は同じ地図を2回とも使用し、情報追随者の場合はそれぞれ異なる地図を用いる。収録された8対話におけるパートナーの組み合わせと地図課題の種類番号は、以下の通りである。

地図番号	情報提供者	vs	情報追随者
--1	A		B
--1	A		C
++2	C		D
++2	C		B
+ -3	B		A

+-3	B	D
-+4	D	C
--+4	D	A

3.4 手続きの統制

実験は、責任者、補助者、録音技師(ホールオペレータ)の三者が担当した。責任者は、実験全体の進行を、補助者は参加者の誘導や控え室の管理にあたった。実験の手順をまとめた簡単なマニュアルを作成し、事前に補助者にわたされている。

3.5 実験参加者の統制

参加者には地図を使った簡単なゲームをすることと、謝礼は後日支払われることが実験参加前に伝えられた。会話の収録が目的であることは知らせない。実験参加者の親近性を統制するため、また参加者が課題に関する情報を交換しないようにするため、以下のような注意が払われた。

- ・控え室は男性ペア、女性ペア別々にし、ペア同志もお互いに話をしないよう管理する。
- ・実験以外では他の参加者と絶対に会話を交わさないように指示する。(とくに実験の内容について他の参加者と話すことを禁じる。)
- ・控え室を勝手に出ないようにする。
- ・参加者の実験室と控え室との間の誘導は補助者が必ず行うこと。(この間も話をしないように補助者は指示すること。)
- ・実験室では他参加者の収録には立ち会わない。

3.6 全体の教示

地図課題が開始されるまで、地図が見えないようにそれはあらかじめ裏返してセットしておく。二人の参加者に対して別々に課題のやり方に関する教示文を読んでもらい、それに関する質問を受ける(責任者が情報提供者、情報追従者別々に誘導しておこなう)。その後、次の教示を与える。

- ・ペンは自由に使ってよい。
- ・ゲームを始めるときと終わったときは、情報提供者がベルを鳴らして合図する。
- ・作業中はなるべくかがみこまないようにする。

教示のあと、録音技師はマイクの位置、音量の調整をおこなう。以上の準備が終わり次第、参加者は地図を表にし、情報提供者がベルを鳴らして始めるよう教示する。課題が終了した時点で情報提供者は再度ベルを鳴らして合図する。

収録後、参加者に基礎アンケートと情報公開に関する承諾書に記入してもらう。

4 収録データ

音声データは、それぞれの話者ごとに独立に2チャンネルでDATに録音された。このデータを16KHz16bitでサンプリングした⁵。各対話における課題の所要時間と転記ファイルサイ

ズをバイト数で示すと以下の通りである。(単位は「分:秒:ミリ秒」) 所要時間は、発話の開始時間から終了時間までの時間数とした⁶。総所要時間は、約 114 分、平均所要時間はおよそ 14 分 14 秒である。

データ番号	所要時間	転記ファイルサイズ
ab	14:27:050	15K
ac	14:16:810	15K
cd	18:34:200	35K
cb	13:42:330	14K
ba	11:55:970	15K
bd	11:00:340	19K
dc	14:32:280	23K
da	15:26:390	26K

5 転記テキストの作成

転記作業をより効率的に進めるために、千葉コーパスで開発された「時間情報作成ツール」および「転記作業支援ツール」を利用し、千葉コーパスの「転記基準」(非公開)におおむね従っておこなわれた。

5.1 転記対象

転記の対象となるのは、音声のうち音韻情報である「発話」および「笑い」である。

5.2 転記基準

発話という単位を厳密に規定することは難しい。今回は、千葉コーパスに従い、音声情報パワーを利用して、客観的な単位としての発話単位(Utterance Unit)を導入した。発話単位とは、「自己発話内の 400 ミリ秒以上の無音区間によって区切られた音声的連続」と定義される(堀内他: 1997)。

さらに句読点に代わる機能となり、発話単位内の分割をおこなう基準として、各発話単位内で 200 ミリ秒以上 400 ミリ秒未満の無音区間について〈 〉の内側にそのミリ秒単位の時間的長さを記入する方法を採用した⁸。発話間の無音区間については記録していないが、発話終了時から開始時までの記録から算出可能である。

また、発話が時間的に重なっている場合は、両話者の発話が 200 ミリ秒以上の重なりを持つ場合を「重複」として、両話者の発話が 200 ミリ秒未満の重なりを持つ場合を「準重複」として、それぞれを区別した。重複は、先行発話と後続発話の同期位置を文字単位で合わせ、重複が開始する場所にそれぞれアスタリスク(*)を記入した。準重複は、その先行発話の最後と後続発話の先頭に+を記入した(堀内他: 1997)。以上のような基準にもとづいて図 2 に挙げるような転記テキストを作成する。

5.3 転記形式

表記はすべてひらがなでおこない、句読点の使用は一切禁止した⁹。延ばす音の表記も必要

最低限におさえた（e. g. 「すたーとちてん」）。

以下、発話時間情報を付与した転記テキスト例をあげる（地図番号は—1）。

00:01:770—00:02:890 G: はじめます

00:04:110—00:05:630 G: あすたーとちてんの

00:07:940—00:09:830 G: みなみがわにたてもものがあるのは

00:10:260—00:10:880 F: はい* あります

00:10:310—00:11:250 G: *わかりますか

00:16:310—00:16:780 G: まず

00:17:430—00:18:710 G: そのたてももの

00:19:330—00:21:890 G: をめざしてすすみますがそのたてももの

00:22:960—00:27:760 G: にしがわをとおっ〈230〉てなんかしてくださいくるまのところ
までなんかしてください

00:32:400—00:32:760 F: えと

00:33:200—00:34:040 F: くるま

00:35:810—00:37:960 G: は〈270〉みなみにみえますか

00:37:600—00:38:450 F: はありません

00:41:030—00:42:220 F: きたにありますけど

00:46:670—00:47:480 F: だいぶはなれてます

00:54:320—00:54:900 G: え

00:57:430—00:58:130 G: じゃあ

00:58:530—01:01:740 G: とそのたてもものくるまのちかくには

01:02:300—01:03:380 G: はたけはみえますか

01:07:980—01:09:090 F: え

01:11:020—01:11:880 F: とあります

01:13:360—01:14:020 G: ありますか+

01:13:400—01:14:170 F: +はいはい

01:18:590—01:22:840 G: じゃあそのたてももの〈320〉はいまきたにみえていますかすたー
とちてんから

01:24:210-01:24:530 F: え

01:25:010-01:26:820 F: といまにしがわにありますけど

01:29:680-01:30:740 F: みなみがわにいくということですか

01:39:400-01:41:100 G: くるまはどっちほうこうにありますか

01:41:650-01:42:910 F: くるまはきたにあります

01:43:040-01:43:140 G: あ

01:49:260-01:50:740 G: でたてものはみなみがわですか

01:51:320-01:51:790 G: にしがわ

01:52:850-01:53:540 F: えといま

01:53:940-01:54:590 F: にしがわです

02:04:340-02:05:480 G: たてもののにしがわ

02:07:650-02:08:730 F: みなみがわにできればいいんすか

02:09:270-02:12:150 G: えと 〈300〉 はたけはどちらがわにみえますか

02:11:980-02:14:750 F: はたけはえと 〈360〉 ひがしがわにあります+

02:14:360-02:15:320 G: +ひがしがわにありますか

図2：転記テキストの例

6 おわりに

本稿では日英語対照研究のための小規模対話コーパスの作成を目指しておこなわれた対話収録の概要と転記作業の内容、転記テキスト例を紹介した。本実験の収録と転記作業は、より信頼できる対話収録とコーパス作成のために、予備実験における問題点を解決することを重点的におこなった。その結果、以下のように改善されたことを報告したい。

1. 録音スタジオを利用できたことにより、2チャンネル音声データを確保することが可能になった。
2. 地図上の目標物のイラストはエジンバラコーパス（名称なし）と同一のものを使用した。また、性別、視認可能条件、親近性条件などの実験デザインはエディンバラコーパスとの相互利用を考慮して、これとほぼ統一できたことは、日英語対照研究のためには有意義なデータを提供することになる¹⁰。

3. 転記作業は、ツールの利用と技術的な協力を多く得て、効率的かつ正確な音声データの書き起こし作業が可能になった。転記テキストの記述は、「読みやすく、かつ直観的に対話の把握が可能な形式」「対話現象を包括的に表現可能な形式」¹¹ を優先する一方で、必要最低限の音声情報を文字化させることに努めた。

自発的で自然な対話コーパスの作成は、対話という複雑な音声コミュニケーション手段をより信頼できる対話データとして正確に記述するための最初の一步であった。エディンバラコーパスとの相互利用のためには、ローマ字表記を一部使用することも今後検討すべきであろう。また、当面の研究目的に応じた転記テキスト記述のための基準（たとえば談話的情報のタグ付けなど）も可能な限り整備していきたい。さらに本稿では触れられなかったが、この対話データは、コーパス利用のためのツールによって、転記テキストを参照しながら音声を再生することが可能となっている。コーパスの実際の分析についてはまたの機会にゆずりたい¹²。

最後に Windows の環境でコーパスを利用した対話分析をおこなうことが可能となったことは、文系対話研究者にとっては大変ありがたいことであり、こうしたことをきっかけとして、今後もデータの共有と対話研究への知見の交流が進むことを望みたい。

註

* 日本語名称なしコーパスの構想は、エディンバラ大学 HCRC で収録、転記されたエディンバラコーパスの未公開部分の一部利用から得られた。コーパスのデザインに関しては、エディンバラ大学理論応用言語学科の Robin Lickley 氏と Jim Miller 氏に貴重な情報提供と助言をいただいた。堀内靖雄氏には日本語名称なしコーパス作成のためのツールおよび情報の提供と転記作業にかかわる詳細な説明をいただいた。対話収録への準備には土屋俊氏、近藤真理子氏、伊藤みか氏および岩敦子氏より多くの実際の助言と資料提供をいただいている。以上の方々に深く感謝いたします。最後になるが、対話の収録と準備にかかわってくれた三重大学人文学部学生（当時）の畑瑞恵氏および転記作業のとりまとめを担当した同大学大学院情報工学研究科の鈴木伸哉氏に深く感謝いたします。また、転記作業に従事した情報工学研究科の学生諸氏の献身的な貢献に対して、この場を借りてお礼を申し上げます。

- 1 コーパスの詳細については Anderson, et al. (1991) を参照。また、HCRC のホームページ上でも公開されている。 <http://hcr.ed.ac.uk/dialogue/maptask.html> コーパスを利用した言語学的研究の一部は、Carletta, Caley, and Isard (1993), Deverell (1994), Miller and Weinert (1998) を参照のこと。
- 2 千葉大コーパスの対話収録と転記作業は 1994 年から 1997 年にかけておこなわれた。コーパスの詳細については青野他 (1994) 堀内他 (1997) を参照。
- 3 予備実験では、地図上の目標物の位置と経路はオリジナルと同じものを使用した。目標物のイラストは異なるものを使用した。今回の実験ではこの点を見直した。
- 4 実験参加者のうち 3 名は三重県出身者、1 名は愛知県出身者である。
- 5 千葉コーパスは 20KHz16bit でサンプリングしている。
- 6 実際的には情報提供者の開始ベルから終了ベルまでの間になる。
- 7 このように定義された発話単位は無音のポーズ (pause) による切れ目におおむね相当すると考えられる。音声言語転記の方法としては、Du Bois, et al. (1993) による 'intonation unit' がある。
- 8 千葉コーパスでは 100 ミリ秒以上の無音区間を採用しているが、今回の転記作業においては作業に困難が生じたためおこなわなかった。
- 9 予備実験では、読みやすさという点を重視して、漢字仮名交じり文による記述を行った。
- 10 ただし、エジンバラコーパスの一部には条件をクリアできていないデータがふくまれている。また、

エジンバラコーパスの音声データはカセットテープのみであるため、2チャンネル音声による電子化は不可能である。この点、現在進行中の転記作業の完了とプロジェクトの報告を待ちたい。

11 小磯他（1995）参照。

12 詳しい内容は堀内他（1997）を参照されたい。

参考文献

- Anderson, A H, M Bader, E G Bard, E Boyle, G Doherty, S Garrod, S Isard, J. Kowtdo, J McAllister, J Miller, C Sotillo, H Thompson, and R Weinert (1991) 'The HCRC Map Task Corpus', *Language and Speech*, 15 (3): 351-66.
- 青野元子, 市川熹, 小磯花絵, 佐藤伸二, 仲真紀子, 土屋俊, 八木健司, 渡部直也, 石崎雅人, 岡田美智男, 鈴木浩之, 中野有紀子, 野中慶子 (1994) 「地図課題コーパス (中間報告)」『情報処理学会音声言語情報処理研究会資料』94-SLP-3-5, pp25-30.
- Carletta, Jean, R Caley, and S Isard (1993) 'A collection of self-repairs from the Map Task Corpus', *Research Paper HCRC/TR - 47* Edinburgh: University of Edinburgh (HCRC Publications)
- Du Bois, John, W Stephan Schuetze-Coburn, Susanna Cumming, Danae Paolino (1993) 'Outline of Discourse Transcription', In Edwards, Jane A and Martin D Lampert, *Talking data: transcription and coding in discourse research* Hillsdale, New Jersey: Lawrence Erlbaum Associates
- Deverell, Jacqui (1994) 'Referring expressions in the Map Task Corpus' *unpublished paper* (Center for Cognitive Science, U of Edinburgh)
- 堀内靖雄, 吉野文, 仲真紀子, 土屋俊, 市川熹 (1997) 「千葉大学地図課題対話コーパスプロジェクト」『千葉大学工学部研究報告』Vol 48. No2. pp33-60.
- 小磯花絵, 堀内靖雄, 佐々木聡, 仲真紀子, 土屋俊, 市川熹, 石崎雅人, 岡田美智男, 鈴木浩之, 中野有紀子 (1995) 「千葉大学課題コーパス作成・利用環境について-1995年度活動報告-」『人工知能学会研究会資料』SIG-SLUD-9503-5. pp23-30.
- Miller, Jim and Regina Weinert (1998) *Spontaneous Spoken Language* Oxford: Clarendon Press
- 吉田悦子 (2001) 「日英語対照研究のための名称なし日本語地図課題対話コーパスの作成-予備実験と転記テキスト-」『人文論叢』第18号. pp173-195.