ハイパーメディア・コーパスの構築と言語教育への応用について

Hypermedia Corpus and Its Application to Language Education

上村隆一
Ryuichi UEMURA

福岡工業大学工学部（福岡市東区和白東3-30-1）

Fukuoka Institute of Technology
3-30-1, Wajirohigashi, Higashi-ku, Fukuoka-shi 811-02
E-mail: uemura@ipc.fit.ac.jp

あらまし：近年、言語研究の分野においても、統計的な分析や資料整理の道具としてのコンピュータ利用が定着してきた。欧米では、大規模な言語データベース（以下コーパス）を用いて、文脈を伴った任意の語彙、語法の出現頻度を調査したり、文学作品、新聞・雑誌記事など特定の分野の言語資料を独自に電子化テキストとして作成し、言語学的な分析結果とともに公開する例が増加している。さらに、最近数年間にインターネットが急成長し、わが国でもその具体的な利用方法が注目されるようになるにつれて、日本から世界へ向けての情報発信の必要性が急速に高まってきた。特に、日本語・日本文化等に関するデータベースを作成し、インターネットを通じて情報を提供することは、わが国に対する国際社会の理解を助け、同時にわれわれ自身が自国の言語・文化を理解し、評価する際の基礎資料としても重要な意味をもつと思われる。

本稿では、著者が研究代表者として開発を進めている日本語会話コーパスの構築プロジェクト（平成7年度文部省科学研究費補助重点領域研究「人文科学とコンピュータ」公募研究（課題番号07207124））について、とくに外国人向け日本語教育への応用の観点から研究成果を中間報告する。

Summary: The object of our joint project started in 1991 is to create an original hyper-media corpus of spoken Japanese. We have collected 'live' spoken data from actual conversation between experts of teaching Japanese as a foreign/second language and native / non-native speakers of Japanese, based upon a testing format known as OPI (Oral Proficiency Interview). The whole contents of experimentation recorded on high precision video tapes and magneto-optical disks were converted to digital video and/or sound data files. The sample version of our corpus is now being transferred to WWW server on our campus (URL: http://corpus. fit.ac.jp) with HTML-tagged texts and is expected to be available over the Internet. This project is authorized and sponsored by Japanese Ministry of Education as one of the 1995 research programs on priority areas (Project No. 07207124). The author is in charge of the research organization (consisting of 5 members including an advisory professor affiliated with U.S. institution) as a whole. It is expected to be completed in 1998.

キーワード：ハイパーメディア、コーパス、日本語、会話分析、インターネット

Keywords: *hypermedia, corpus, Japanese, conversational analysis, Internet*

1．研究経過

本研究は、日本語母国語話者(以下NS)と非母国語話者(以下NNS)の現実発話に含まれる言い誤りの類型を比較分析することを目的として、1991年度より開始した試験研究の延長線上にある。研究当初から、分析対象となる一次言語資料の絶対量不足を痛感したため、われわれはまず、インタビュー実験形式によ

る会話データの収集と、それに基づくコーパスの構築作業から開始することにした。

平成７年度は主としてNNSのデータ収集を行うことになり、７月に東京都内の民間日本語学校と国際基督教大学、10月に米国のプリンストン大学においてそれぞれインタビュー実験を実施し、約70名分の会話データを得た。（被験者の内訳は**表１**のとおり。）

表１　インタビュー実験被験者の内訳

国籍	米国 *26*　韓国 *25*　中国 *5*　日本 *3* ロシア・オーストラリア 各 *2* ドイツ・オーストリア・タイ 各 *1*
性別	男 *29*　女 *37*
年齢層	20代 *58*　30代 *7*　40代以上 *1*

日本語学校の被験者は１名を除いて全員が韓国籍で、年令は全員20歳代であった。大半が就学生で、日本語学習歴、日本滞在期間、生活環境も似通っている極めて均質のグループであった。次に、国際基督教大学では夏期日本語講座の受講生を被験者としたため、被験者の国籍は多岐にわたり、母語や言語背景も多様であった。この傾向はプリンストン大学についても同様で、欧米系とアジア系がほぼ同数であった。年齢層、男女比なども適当に分散していたと思われる。

会話データの収録に際しては、事前にこの研究の趣旨とリスクに関して説明を行い、第１回目は口頭での承諾、２回目以降は同意書への署名いう方法をとった。確認事項は次の通りである。

1. この会話データ収録は15～30分の日本語によるインタビューである。
2. このインタビューは録音、録画される。
3. このデータは研究目的以外には使用されない。
4. このデータはインターネット上で公開される。
5. 4に関しては現時点で予測不可能な犯罪行為を含むリスクについては、研究者および実験者は一切責任を負わない。

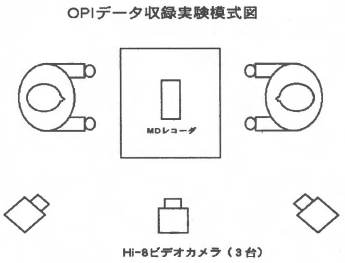
個人データのインターネット上での公開に伴うリスクについては、現時点では不透明な部分が多く、法律的な安全策も未だできていない。このような状況の下では、今後も事前に被験者に十分な説明を行うことは研究者のすべき最低限の配慮であろう。その上で被験者の了解を明確な形で得ておくことが研究計画の成功には不可欠であろうと考える。(注1)

２．実験方法とデータ処理

会話データの収録形式としては、NNSの会話能力判定方法として知られるOPI(Oral Proficiency Interview)を採用した。これは会話モードとロールプレイモードの二つの要素によって構成されるインタビュー形式のテストであって、本来は外国語学習者の口頭表現能力を総合的に評価することを目的としたものである。従って、他の多くの会話体データの収集方法と異なり、実験者はできるだけ被験者に自発的に多く話をさせるように配慮した。

データ記録媒体については、画像データを高画質8ミリビデオテープに、音声データを光磁気ディスク（MD,ディジタル録音専用メディア）にそれぞれ収録した。収録時間は被験者１人につき20-30分程度である。（**図１**参照）

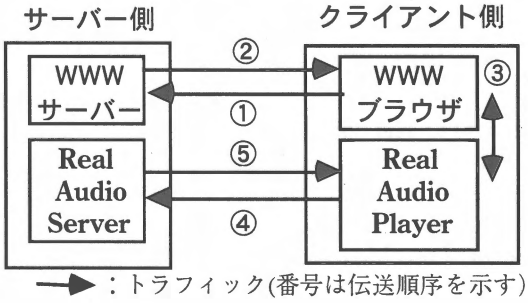
図１　実験機器類セッティング



現在までに、音声データからテキストデータへの書き起こし作業の一部（約10名分）が完成し、圧縮した音声データとともにインターネットのWWWサーバー(http://corpus.fit.ac.jp)およびFTPサーバー上で順次公開している。

特に、音声データの公開方法に関しては、最近インターネット上でのオン・デマンド型音声サーバー技術として注目されているReal Audio Serverをいち早く導入し、本年９月から実験的にインタビューの内容の一部（フリートーキングとロールプレイについて各々約20人分）を転写テキストと一体化した形で提供している。（**図２**参照）

図2　オン・デマンド音声サーバ概念図



ディジタル変換された画像データ（動画）は被験者1人当り200～300MB（圧縮ファイル）に達するので、現時点ではインターネット上では公開せず、追記型光ディスク（CD-R）に保存している。また、このコーパス作成作業と同時に、繋ぎ語や代名詞類の言語学的分析を同時に進めており、それらの研究成果の一部は、情報処理関係の学会で発表し、さらに言語学関連の研究論文集等で公刊している。

3．本研究の特色

本研究プロジェクトにおいて実現をめざす日本語会話コーパスは、従来のテキスト・データベースと異なり、文字・画像（動画）・音声データを統一された使用環境(GUI)で同時に利用可能にする、いわゆるマルチメディア型データベースである。われわれのコーパスの主な特徴は、

①分析対象が話し言葉である場合、文字化しにくい音調、強勢、ポーズなどの諸特徴をそのまま音声情報の形で提供できる。その結果、テキストデータに特殊な音調記号等を付与する必要がなくなる。

②画像（動画）データを提供することにより、非言語情報（身ぶり、手ぶり、顔の表情など）をテキストと同時に利用できるようになり、会話の状況、話者の特徴、周囲の雰囲気などを分析の手がかりにすることができる。

③画像・音声データ自身をディジタル化することにより、ランダム・アクセスが可能になるので、テキスト検索に加えて、画像・音声データの検索等を容易に実行できる、などである。

上記のコーパスの特性を十分に活用することにより、これまで研究対象から事実上除外されてきた非言語情報を含む会話分析が可能になる。また、従来のコーパスのように、転写記号等に関する専門知識を必要としないので、研究用資料としてだけでなく、外国人向けの日本語教育の資料・教材としても十分に活用できるであろうと思われる。

4．日本語教育からみた本研究の意義

本研究プロジェクトが日本語教育の現場に与えるインパクトは限りない。その一つを挙げれば、本年度実施した2回の会話データをみるだけでも、第二言語の習得に関わる諸要因についての多くの示唆を読み取ることができるだろう。（転写テキストの具体例は本稿末尾の**サンプル・データ**を参照。）さらに数多くのデータが得られれば、個々の要因についての研究も可能になることが期待される。

さらに会話データの中のロールプレイモードのヴァリエーションは、社会的文化的に適切な言語使用を教師がどのように指導すべきなのかという問題に対する解決の糸口を与えれくれる。次年度以降に予定している、NSの会話データ収集が進めば、現在多くの日本語教師が持っている社会的・文化的に適切な言語使用のイメージの見直しを迫られることになるかもしれない。(注2)

5．今後の研究計画

1996年度は本年度に引き続いて、会話データの収集、テキスト転写作業を行う。また、情報処理、日本語教育関連の学会において、本研究の第2次中間報告を発表する。実験データの収録作業については、同年度内に次の2点を実行する。

1）米国在住の日本人研究協力者（牧野成一プリンストン大学教授）の指導の下に、国内と同一の実験条件で現地在住のNS,NNSそれぞれの会話データを収録。

2）研究分担者と国内2大学の協力の下に、主としてNSの会話データを収録。

音声データについては、オーディオテープ(DATを含む）やMDに録音した場合、データベース作成時に膨大な量の変換作業を必要とするため、別途ノートブック型パソコンと16ビットサンプリング可能な音声入力・編集ソフトを用いて直接ディジタル録音を試みる。このことにより、書き起こし（転写）作業から音声データベース部分の構築に至る作業工程を大幅に短縮することができる。

実験で収録したデータは、日本語教育関係の研究分担者が大学院生等に委託した形で転写作業を行った後、責任を持って校閲する。転写作業にあたっては、16ビットサンプリングによりパソコン上で直接ディジタル録音したファイルから、ディジタル録音・編集機を用いて適宜分割・再編集した音声データを用いる。

1997年度は研究班をコーパス作成班とコーパス分析班に分けて、適宜協力しながら研究計画を遂行する。まず、「作成班」は前年度内に収集した音声データと転写テキストデータについて、任意の文字列から当該個所の音声データを検索するプログラムの開発を試みる。特に、時系列データを扱う検索用言語としては、HyTime処理系等を参考にしながら、SGML／DTDの拡張を考える。同時に、ディジタル動画データ(MPEG形式）の検索方法も検討する。なお、コーパス本体は上記のインターネット上で提供するネットワーク版と、CD-ROMで提供するスタンドアロン版の2種類を作成する予定である。また、CD-ROM版

の検索ソフトウェアの更新及び追加情報の提供等はすべてインターネットを利用し、国内外の言語研究者および日本語教育関係者に公開する。　次に、「分析班」は言語学と日本語教育の立場から、NSとNNSの発話内容を比較検討し、段落形成、連接名詞、代名詞類、繋ぎ語等の各トピックについて会話分析を試みる。データの蓄積が一定レベルに達した後、当初の研究目的であった「言い誤り」の類型に関する分析を開始する。年度内に本研究プロジェクトの最終報告を行うが、論文とデータは通常の印刷物に加えて、電子化テキストの形式で作成し、インターネット上のFTPサーバーからも利用可能にする。

6．おわりに

コーパスを用いた言語分析は、欧米ではすでに確立した研究手法であり、分析対象も文語体のテキストにとどまらず、会話内容を転写したデータから成る口語体テキストにまで及んでいる。日本でも、最近コーパスの重要性が認識され、欧米のLOB,London-Lund, Brown, BNC等の大規模コーパスを利用した英語の文法・語法などの研究例が増加しているが、日本語コーパスについては、いまだ欧米ほど確立した大規模なものがなく、まとまった研究成果も報告されていない。従って、われわれの研究は、日米間にまたがる大規模な日本語会話コーパスの構築プロジェクトとしては前例のないものであり、インターネット環境およびマルチメディア型データベースを使用した言語研究としても、先駆的な役割を果たすもの、といえる。

*注）1,2ともに共同研究者の村野良子氏（国際基督教大学・日本語教育）の指摘による。

参考文献

Armstrong, S.(Ed.) : *Using Large Corpora.* MIT Press. (1994).

Buck, K. (Ed.) *The ACTFL Oral Proficiency Interview Tester Training Manual.* The American Council on the Teaching of Foreign Languages. (1989).

Uemura, R. : "Hypermedia Corpus Project of Japanese Conversation - Interim Report," *Language and Information Processing* **6,** pp. 1-5. Fukuoka Institute of Technology. (1995).

上村隆一「コーパスによる日本語会話分析ー指示詞の使用について」小泉保教授古稀記念論文集『言語学の展望』 pp.93-105. 大学書林. (印刷中).

（参考）サンプル・データ

（1：は実験者、2：は被験者を示す。カッコ内は相づち、／は長いポーズを表す。）

A. 会話モード（フリー・トーキング）

1： ああーそうですか。（2：ええ）最近の日本のもん、ま、色々問題を抱えていると（2：そうですね）思いますけれどもね、そのー、どうですか。まあ、まよく聞かれることだと思うんですけど、オウム真理教、あたりがね、そのどうして日本の社会に、ああいう今までにないようなね、（2：ええ）規模の、おー、ほんとに、なんて言うか、極悪、（2：うん）非道なね、（2：うん）ああいう、一つの、そのグループが出来たっていうのは、どういうことでしょうね。（2：すー）なんか歴史的な観察、ありますか。
2： ええ。まあ歴史的な観察は／まあ歴史にも、そんなグループが、ないんですけど、でも似てるグループがあると思いますよ。（1：うん）日本の社会は、アメリカ人の考えは、日本の社会は平和で、えーといつも、目上の人に従う社会ですね。でも、ま、歴史の立場からみるとー、ま、色々な、えーとほんとに、強く、反対した人が、いましたねえ。
1： しかしだけど、オウム真理教とは（2：オウム真理教と）ちょっと、比較するとちょっとまずいんじゃ（2：ちょっと違いますねえ）ないんですか。

B. ロールプレイ・モード

2： あのー／。あのー／。うん。
1： うん、だあれー、お姉ちゃん？
2： あ、あのー、／うん、／あのー、あなたの名前てー、／山田ー／ていうのかなあ？
1： うん山田ーだけど。ぼく山田一郎。
2： あ、そう。（1：うん）じゃあ、あの、お父さん、／ちょっといらっしゃるー？
1： うん、お父ちゃんいないよ。
2： ああ、あ、そう。じゃあ、あのー、うんー、どうしようかなあ。わたしー、お父さんにちょっとー、あのー、お父さんいらしたら、ちょっとー、うん、お会いしたいと思うけどー、お父さん何時ごろ、お帰りになるの？
1： お父ちゃんねー、なんか、明日帰ってくるっていってたよ。
2： あ、そっかー。じゃあー、お母さんは？いらっしゃー、らないの？
1： 今ねえ、お母ちゃんねえ、今、買い物行ったみたい。
2： そっか。じゃあ、あのー、ここでちょっと、待っててもいいかなあ？
1： うーん、たぶん、うん待っていていよ。うん、なんか、一緒に遊んでくれる？
2： あ、いいよ。
1： あの今パソコンやってんだけど。
2： あ、じゃあ、あの、パソコンやりながら、ちょっと、お母さん、待つわ。

Hypermedia Corpus of Spoken Japanese

PREVIOUS

MENU

 T.K.(Japanese)  H.L.(American)

 K.S.(Korean)

 C.Y.(Korean)  B.P.(American)

 M.S.(Austrian)

 D.N.(German)  M.I.(Russian)

Click on any picture icon to access each data.

Please feel free to send your comments to:
uemura@ipc.fit.ac.jp

CORPUS SAMPLE (Interview Setting)

MENU

Click on any picture icon to hear each sound data.

Free Talking

Role Play

Free Talking

１：はい、こんにちは。
２：こんにちは。
１：あの、私の名前は牧野と申しますが。
２：牧野先生ですか。
１：はい、牧野です。そちらは、お名前は。
２：ブライアン・プラットです。
１：あ、じゃ、ま、ブライアンさんて呼んでいいですか。
２：う、ブライアンでいいです。
１：ああ、そうですか。
２：はい。
１：ブライアンさんは、アメリカ人ですか。
２：ええ、そうです。
１：ああ、どこ、アメリカのどこからいらっしゃったんですか。
２：えと、出身はウエスト・バージニア州ですが、（１：あ）あの、今イリノイ大学の大学院生です。
１：あ、そうですか。ああ、私はイリノイ大学で前教えていたんですけれども。
２：ああ、わかりますよ。
１：あ、知ってますか？ああ、そうですか。ああ、そうですか。ウエスト・バージニアっていうのは小さな州（２：そうですね）ですよね。ええ。私はあすこに一度行ったことがあるんですけれども。
２：ああ、そうですか。ありますか。
１：うーん、まあ、行ったことがない人にどういうふうな州だって説明できますかね。
２：そうですね。
１：ちょっと説明してみてください。
２：まあ、アメリカの文化からちょっと離れていると思いますね。（１：うん）アメリ、ウエスト・バージニア州で。（１：うん）ええと、まあ、丘がたくさんあるし、（１：うん）ええと、離れている町が多いんですね。（１：うん）だから、その、離れている町に住んでいる人は、（１：うん）まあ、世界のことか、アメリカのこと、ま、よく知らない人が多いんですね。だから、えーと、ま、外国人に会ったことがない人もいるし、（１：うーん）えーと、まあ、世界のことを全然構わない人もいるし、ちょっと、も、面白いところだと思いますね。

Role Play