マルチメディア型言語データベースの構築とその応用について

上村隆一

Project: Hypermedia Corpus and Its Application

Ryuichi Uemura

Our joint research project of creating an original hypermedia corpus of spoken Japanese began in 1991 as a pilot study for error analyses in discourse. We have collected 'live' spoken data from actual conversation between experts of TJFL (Teaching Japanese as a Foreign Language) and several native and non-native speakers (or learners) of Japanese. The experiments were executed in the forms of free talking plus role plays, which were based upon an established testing format known as OPI (Oral Proficency Interview). The whole contents of OPI experimentation were first recorded by camcorders on several high precision video tapes, and then converted into digital video files and/or digital sound data files, using DTV hardware and software. So far, the accumulated data (compressed files) have amounted to more than 1GB, which include movie clips, sound files, and full texts transcribed from the original digital sound data. The process of text transcription and its link to video and audio data is still under way. Our research has also been motivated by a new concept of *global network*. Recently, a sharp rise of interest in communicating ideas and sharing resources via Internet seems to be accelerating the move toward building up distributed databases, accessible to any user connected to the worldwide network, e.g. WWW, Gopher and WAIS. The sample movie data of our corpus project is now being transferred to WWW server (http://www.fit.ac.jp) along with HTML-based tagged texts and is expected to be available on the Internet.

1. はじめに

近年、わが国でも学術用データベース作成と共同利用環境づくりの必要性に対する認識が深まってきたように思われる。特に、技術情報の蓄積や論文検索を主体とした、どちらかというと理科系中心の利用環境から、文学作品、歴史資料、文化財等の文科系寄りの利用環境へと広がりを見せてきたのが最近の特徴である。また、欧米の研究資料を受動的に利用するだけでなく、独自のデータベースを作成し、インターネットを通じて国際的な共同利用を可能にしようとする能動的な気運も徐々に高まってきている。とりわけ、日本語・日本文化等に関するデータベース作成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。本研究も、そうした「グローバル・ネットワーク」の視点に立ち、インターネット上で利用可能な、動画・音声・文字情報を含む日本語会話データベースを構築するプロジェクトである。研究成果の発表についても、通常の論文形式に加えて国際標準化されたタグ付き電子化テキストを所属機関のftpサーバーに随時登録、公開していく予定である。

2．研究経過

本研究プロジェクトは、日本語の談話構造に関して、日本語母国語話者(以下NS)と非母国語話者(以下NNS)の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年度より試験研究が始められた。日本語の会話分析は、まだ体系的な研究が少なく、とくに分析対象としての一次言語資料（以下コーパス）の絶対量が少ない。

そこで、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。これまでに、２大学（国際基督教大学、成蹊大学）の教職員および学生の協力によって、NS, NNS計15人の会話データをビデオテープに収録し（被験者１人につき30分程度）、テキスト転写、画像・音声のディジタル化作業を経て試作版コーパスを作成中である。現時点で、転写されたテキストデータはNSの部分だけで被験者１人当り約100KB（ただし注釈等を含む）、ディジタル画像データ（動画、１秒間３０コマ再生）は被験者１人当り200～300MB（圧縮ファイル形式、圧縮比１：８程度）に達しており、追記型光ディスク（CD-Recordable）に保存している。また、このコーパス作成と並行して、繋ぎ語や指示詞等の使用実態の分析（後者については本予稿末尾の実例を参照）を同時に進めており、それらの研究成果の一部は、第１次中間報告の形ですでに国際学会(*2nd Princeton Japanese Pedagogy Workshop*)等で共同発表してきた。さらに、筆者単独の研究としては、所属校の言語情報工学研究所の補助を受けて同コーパスの検索機能等の充実、多角的な分析作業を継続中であり、同研究所紀要に所収予定の論文およびその他公刊予定の論文集によって、その研究成果を報告するべく準備を進めている。

3．研究内容の特色

本研究プロジェクトにおいて作成する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、以下の通りである。

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. 今後の研究計画

これまでの試験研究の段階では、被験者の確保、実験条件の整備等の諸問題があり、結果的にNNSのデータがNSに比較して相対的に少なくなった。また、人種的、年齢的な偏り（大部分が欧米系で20歳代の留学生）および日本語能力の格差も相当にあり、標本分布の上でバランスを欠いた。さらに、NS,NNSの両者について、言語行動の国際比較という観点から、日本国内だけでなく、外国在住のNSおよびNNSにまで分析対象を拡大して比較検討することが重要である、と思われる。

そこで、平成7年度以降、以下の順序で追加データの収集、コーパス構築作業を行う。

1. 日本語教育関係者（日本語会話能力テスター有資格者）と同出版社（株式会社アルク）および数校の日本語学校の協力を得て、非欧米系（アジア、中近東などの出身者）NNSの会話データを収録。
2. 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。
3. 日本在住の欧米系NNSおよび日本国内のNSのデータ追加収録については、国際基督教大学で、村野の責任において実施。
4. 収集した言語データ（ビデオテープに収録）について、ある程度の取捨選択を行った後、集中的にテキスト転写、ビデオ部分のディジタル化(MPEG又はQuickTimeMovie)を実施し、最終的にNS, NNS合わせて50人分の画像・音声データのディジタル化と編集作業を完成させる。

完成したコーパスはCD-ROM上で利用可能な形にし、言語学、日本語教育および情報処理関係の学会で談話分析の結果とともに研究発表する。また、テキストデータとビデオデータ（ディジタル・ムービー）はそれぞれインターネット上の分散型データベースサーバー（WWW , ftpなど）に登録し、国内外の言語研究者および日本語教育者に公開する。

5. 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-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、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。

（参考）コーパス作成方法とデータ検索の実例

会話データの収集方法については、日本語会話能力検定の手法として知られるOPI(Oral Proficiency Interview)に準拠し、インタビュー形式の実験を行う。被験者の数はNS, NNS各25名程度である。実験者と被験者それぞれの視点の延長上および実験者と被験者の中間

（実験状況を確認するための参照画面用）の計3カ所に高画質8ミリビデオカメラを据え、OPI実験の模様をビデオテープに記録する。

実験で収録したデータは、各大学の研究者・日本語教育関係者が分担して転写作業を行う。転写作業にあたっては、原則としてアナログ音声をディジタル変換したデータファイルを適宜分割し、音声編集可能なノート型パソコン（可搬性を考慮）に移植し、ワープロソフトを用いて行う。

転写作業完了以後のコーパス作成手順を示すと、およそ下記のようになる。

①実験データを収録したビデオテープから画像・音声データをディジタル形式に変換
②データベース作成支援ソフトを用いて、テキスト・画像・音声データをリンク
③ディジタル化された画像・音声データを適当な単位に分割し、データ名を付与
④ページ単位で画像と音声を再生するための注釈ボタンをそれぞれ設定
⑤CDライタを用いて追記型光ディスク(CD-R）に全ファイルを保存。

最後に、試作版コーパスを用いた代名詞類の検索例を紹介する。文語体の場合に比べて、現実の発話においては、代名詞が単なる先行名詞の照応表現以上の機能を果たしていることが以下の実例によっても明らかである。なお、会話中の１は実験者、２は被験者を示す。また、カッコ内は聞き手のあいづち、下線部が代名詞（ここでは「あれ」）、*印で挟まれた部分が指示対象の名詞（句）を示している。

（後方照応による記憶の想起）
１：そうですね。あの、ご出身もー、あれですか？　あのー、*東京都*ですか？
２：ええ、東京です。
１：それは単に、あのー、あれですか、保谷から国分寺高校っていうのは、単に、その、*電車通学に憧れて[ということ]*？
（聞き手による指示対象の補完）
２：[夏休みは]取れなくないかもしれないんですけど（１：うーん。）それほどにまだ、（１：ああ。）あの、どこへ行くっていう、あれもないですから。
１：ああ、そうですか。（２：ええ。）全然*予定*は立ててないっていう、うーん。
（話者交替による先行詞の引き継ぎ）
２：それで、特別にあと、プロジェクトをやったときには（１：うん。）一千万ぐらいはもらって、何かこういう*システム*を作ります。
１：え、じゃあ、すごい、あれですね。
２：そうですね。

参考文献

上村隆一、田吹昌俊　(1994)「ハイパーメディアコーパスと日本語会話分析」*Proceedings of 2nd Princeton Japanese Pedagogy Workshop* pp. 184-200.

福岡工業大学工学部助教授
Associate Professor
Faculty of Engineering, Fukuoka Institute of Technology