ｍｉｃ－Ｊコーパスの公開について
「外国人へのインタビュー篇[1]」「日本人へのインタビュー篇[2]」

西郡 仁朗[*]　崔 文姫[**]　磯野英治[***]

1. はじめに

現在の言語研究，言語教育研究において、実証性や客観性の確保のためにコーパスの利用を要件とするものが増加し、集積したデータを無償または安価で公開する学術的な機関・グループも多くなって来ている。これまでは書き言葉に関するコーパスが多く、代表的なものとしては『青空文庫』,『新潮文庫の１００冊』，新聞各社の新聞記事データベース,『国会会議録検索システム』などがネット上、またはデジタル素材として公開され利用されてきた。最近では話し言葉に関しても公開が始まっており、国立国語研究所の『日本語話し言葉コーパス』がその代表的な例であろう。社会言語学や第二言語習得論などでの利用や日本語教育学への応用を考えると、日本人の話し言葉の文字起こしされたデータだけでなく、外国人による日本語の話し言葉や日本語使用場面でのデータの集積も必要であり、現に宇佐美(2003)の BTSJ(Basic Transcription for Japanese)を用いた一連のデータ公開や、同じく BTSJ を用いた西郡(2002)による方法論とデータの公開では、外国人を含めた日本語会話のデータが含まれている。さらに日本語の OPI (Oral Proficiency Interview)の知見を利用した KY コーパス（鎌田, 2006）のデータも公開されており、このデータをもとにした研究も行われている。

日本語教育学の研究と教育への応用を考えると、日本人同士、日本人と外国人、外国人同士のデータの集積と公開は不可欠のものであり、さらに、パラ言

---

[1] http://japanese.human.metro-u.ac.jp/mic-j/mic-J_corpus_FI/index.html
[2] http://japanese.human.metro-u.ac.jp/mic-j/mic-J_corpus_JI/index.html

[*] 首都大学東京 Tokyo Metropolitan University
[**] 国士舘大学 Kokushikan University
[***] 韓国・中央大学校 Chung-ang University, Korea

語や非言語行動の分析を考えると、文字起こしのデータだけでなく、音声・動画のデータも同時に収録することが望ましい。既述の宇佐美の BTSJ の WEB サイト[3]では、動画データの公開も始まっている。

本稿では、筆者らが制作した２つの会話資料「外国人へのインタビュー篇」と「日本人へのインタビュー篇」の『mic-J コーパス』としての公開について報告する。

これまで、大学内部のデータとして利用され、研究対象ともなって来た（崔, 2009a; 崔, 2009b; 磯野,2009a; 磯野, 2009b; 磯野, 2009c)。しかし、データの利用可能性が大きく、また、公的助成を受けていることもあり[4]、学内外・国内外の利用に供することとした。

公開コーパスにおいては、BTSJ による文字起こしデータ、音声を含む動画の資料も掲載されており、全データが閲覧だけでなくダウンロードが可能となっている。

2. 公開コーパスの作成手順

今回公開するデータは「外国人へのインタビュー篇」と「日本人へのインタビュー篇」の２種類である。両者は収録時期も内容も異なるものであるが、作成した手続きはほぼ同じであるので本稿でまとめて記すこととする。

作成は、この分野での研鑽を積む為の大学院のコースワークの一環であったが[5]、下記の一次文字起こし等では一部有給作業として行われたものもある。

作成においては、BTSJ データと動画を含めた一般に公開されるコーパス作成であることを関係者が同意し、コーパスの意義や BTSJ によるデータの保存法などについて技術的な面も含めた研修を行った。また、どのような会話内容とするかが話し合われた結果、インタビューの形式をとること、インタビュワ

---

[3] http://www.tufs.ac.jp/ts/personal/usamiken/index.htm

[4] 本プロジェクトは一部、東京都アジア人材育成基金の助成により行われた。

[5]首都大学東京大学院人文科学研究科日本語教育学教室「日本語教育学研究・特論　談話と音声教育１」で「外国人へのインタビュー篇」は 2007 年度前期,「外国人へのインタビュー篇」は 2008 年度前期に行われた。

ーは３０代後半のプロの人間に依頼して円滑で出来るだけ自然な会話の流れとすること、インタビュイー（インタビュー対象となる被験者）はインタビュワーよりも年下で、日本人・外国人それぞれ２０代の者で性別のバランスをとること、外国人については母語と日本語運用レベルについてもバランスをとることとした。ただし、外国人の被験者確保が困難な面があり、年齢については外国人で３０代の者が３名、１０代の者が２名含まれることとなった。また、日本語運用レベルについては『日本語能力試験出題基準』(1994) に示されている基準を参考に２級程度のレベルを中級に、１級以上のレベルを上級とし、フェイスシートの情報や面談などから総合的に判断して他の属性とのバランスをとることとした（表１および表３参照)。

表１　コーパス作成の手順

	外国人へのインタビュー篇	日本人へのインタビュー篇
撮影日時	2007 年 6 月 1 日, 2 日, 9 日	2008 年 6 月 21 日 22 日
撮影場所	首都大学東京 飯田橋キャンパス 31 教室および事務室	首都大学東京南大沢キャンパス 国際交流会館 中会議室
インタビュイー	外国人２０代（一部 30 代 10 代）計２０名 ５母語（中国語・韓国語・タイ語・インドネシア語・英語) ×性別（男・女) ×日本語レベル（上級・中級) （表３参照)	日本人２０代男女計２４名 男性 12 名 （大学院生２名、学部生 10 名) 女性 12 名 （社会人２名、学部生 10 名)
インタビュワー	杉山裕子（シグマ・セブン）	杉山裕子（シグマ・セブン）
撮影	崔文姫　西郡仁朗　趙恩英　王威　魯菲　蹇敏　磯野英治 ＜撮影協力＞㈱ジャパンライム	磯野英治　西郡仁朗　劉永亮 柴田沙矢香　王艶　閔竣泓 ＜撮影協力＞㈱ジャパンライム
一次文字起こし	崔文姫　趙恩英　王威　魯菲 蹇敏	磯野英治　柴田沙矢香 渡邊千佳子　小玉博昭
二次文字起こし	崔文姫	磯野英治
発話文認定	崔文姫　西郡仁朗	磯野英治　西郡仁朗

表２　インタビューでの質問内容

外国人へのインタビュー篇	日本人へのインタビュー篇
名前と簡単な自己紹介。 日本に来て何年か？ 日本料理を食べるか？ よく食べますか？ 特に好きな日本の食べ物があるか？ その料理をどこで食べるか？ あなたの国に日本料理のレストランがあるか？ そこにも行ったことがあるか？ あなたに国で人気のある日本料理は何か？ あなたの国の日本料理のレストランの値段はどうか？ 日本の食べ物で嫌いな物はあるか? 日本の食習慣の中で気になる点はあるか？ 日本料理、日本の食べ物についてどんなイメージを持つか？	名前と簡単な自己紹介。 アルバイトをしているか／したことがあるか？ またその内容は？ アルバイトのとき何か困ったことや大変だったことがあったか？ アルバイトをしていて良かったことや嬉しかったことがあったか？ 将来どのような仕事に就きたいか？ あなたの学校の好きなところは何か？ 学校に不満があるとしたら何か？ 休みの日には何をすることが多いか？ これまでの旅行で印象に残っていることは何か？ 次に旅をするならどこへ行きたいか？ 好きな映画、面白かった映画は何か？ 外国人の友だちがいるか？ いるならどんな話をするか？ 外国人に日本を案内するとしたらどんなところへ連れて行きたいか？ 今年、チャレンジしたいことは何か？ 一生のうちに一度はしてみたいことは何か？

インタビューの内容については、外国人については「食生活」をテーマにすること、日本人についてはアルバイトや現在の夢などの一般的な質問とし、それぞれ項目を列挙してインタビューを行うこととした（表２参照）。各項目についての質問は出来るだけ自然な話し方でプロのインタビュワーが質問することにしていたので、相づちなど聞き手としての反応、確認や追加の質問・発展質問はインタビュワーの技術に任せ、設定された質問内容から大きく逸脱しない限り、自由に進行できることとした。

本コーパスの外国人へのインタビュー篇を研究目的で利用する場合、インタビュイーについての情報が重要になると思われるので、表３に詳述する。

表３　インタビュイーについての背景情報

インタビュイー	性別	年齢	国籍	判定レベル	滞日期間	学習時間	能力試験
1.KRMA	男	29	韓国	上級	２年２ヶ月	日本語学校１年半（上級クラス）修了	１級
2.KRFA	女	20	韓国	上級	11 ヶ月	日本語学校上級クラス在籍＋高校で 480 時間	
3.KRMI	男	25	韓国	中級	11 ヶ月	日本語学校中級クラス在籍	
4.KRFI	女	26	韓国	中級	６ヶ月	日本語学校中級クラス在籍＋韓国の専門校で 320 時間	２級
5.CHMA	男	24	中国	上級	１年８ヶ月	日本語学校１年半（上級クラス）修了	１級
6.CHFA	女	30	中国	上級	５年２ヶ月	日本語学校２年修了＋日本の大学卒業	１級
7.CHMI	男	26	中国	中級	１年８ヶ月	180 時間（中国の語学学校）	２級
8.CHFI	女	25	中国	中級	２年	日本語学校１年修了＋大学授業 90 時間	２級
9.INMA	男	20	インドネシア	上級	１年２ヶ月	本国国際交流基金で１年修了＋大学授業 90 時間	
10.INFA	女	24	インドネシア	上級	４年２ヶ月	日本語学校１年修了＋大学等での学習 200 時間＋５ヶ月日本に交換留学	１級
11.INMI	男	18	インドネシア	中級	８ヶ月	日本語学校 204 時間＋大学授業 72 時間	３級
12.INFI	女	29	インドネシア	中級	２年１ヶ月	日本語学校中級クラス在籍＋ユネスコ 140 時間	２級
13.THMA	男	20	タイ	上級	２年２ヶ月	日本語学校１年修了＋大学授業 90 時間	１級
14THFA	女	30	タイ	上級	２年２ヶ月	大学別科 576 時間＋大学授業 1200 時間	１級
15.THMI	男	21	タイ	中級	９ヶ月	日本語学校中級クラス在籍＋その他約 240 時間	
16.THFI	女	19	タイ	中級	９ヶ月	日本語学校中級クラス在籍	
17.EGMA	男	27	カナダ	上級	３年３ヶ月	日本語学校上級クラス在籍＋他に約 240 時間	２級
18.EGFA	女	24	イギリス	上級	11 ヶ月	日本語学校上級クラス在籍＋本国で約 100 時間	
19.EGMI	男	37	アメリカ	中級	10 ヶ月	日本語学校中級クラス在籍＋本国で約 200 時間	３級
20.EGFI	女	24	イギリス	中級	８ヶ月	日本語学校中級クラス在籍	

インタビュイー欄ではじめの２文字は母語を示す（KR:韓国語，CH:北京語，IN:インドネシア語，TH：タイ語，EG:英語）。能力試験の欄は取得済みの資格。表は崔(2009a)による。

実際の撮影においては、①コーパス作成の趣旨とデータ公開についての説明，②フェイスシートの記入と日本語学習背景などについての面談，③撮影，④撮影に関するインタビュイーからのコメント，⑤素材公開への同意書作成，の順で行われた。同意書では、この素材が動画と音声を含めインターネットで公開されることが明記されている。また、インタビュイーとインタビュワーには謝礼が支払われている。

文字起こしに関しては、宇佐美(2003)の BTSJ の方法に従っており、一次文字起こし，そのチェックを含めた二次文字起こし、発話文をどこで区切るかに関する複数による部分チェックと Cohen's κの算出と一致度の確認（参照：西郡，2002）の順で進行した。

3. サイトの構造

上記のような手順でデータ化された内容は、サイト上で図 1 に示す構造で閲覧可能となっている。また、実際の WEB ページの例を図 2 に示す。

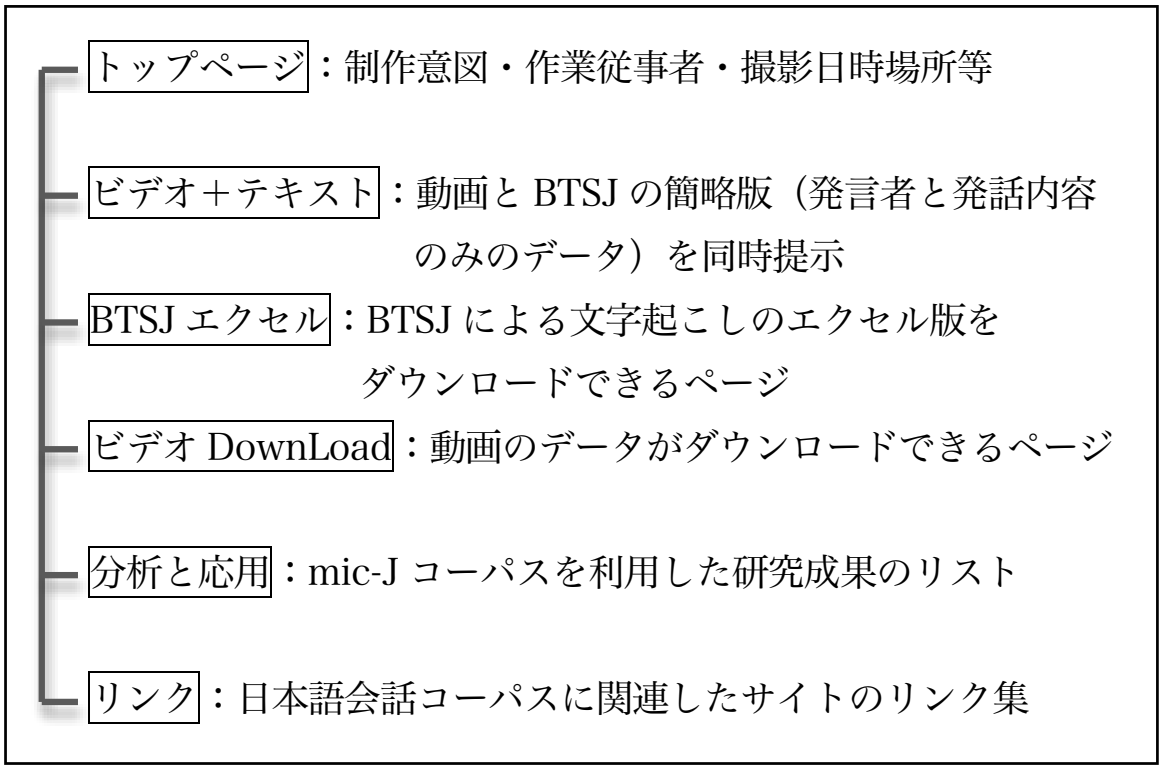

図 1　サイトの構造

（「外国人へのインタビュー篇」「日本人へのインタビュー篇」共通）

図２　動画とテキストデータを提示する WEB ページの例

3. おわりに

以上、集積・整理したコーパスのデータの公開について記して来た。今後の多くの研究者によるデータ集積と公開をするが、いくつか困難な点があるので記しておきたい。

まず、良質な音声・画像の収録は、専門的な技能がないと難しい。最近は電子光学機器の進歩により民生品でも相当の良質な画像が撮影できるようになってきたが、それでも、カメラアングルやカメラワーク、ライティングなどに注意しないと非言語行動などの記録が粗雑になる可能性がある。音声の収録はさらに注意が必要で、カメラ内蔵のマイクでは雑音の少ない環境でもインタビュイーから２メートル以上離れると正確な収録は難しい。やや高額だがワイヤレ

スピンマイクのセットが必要になる。

文字起こしの方法については統一された方法がない点が問題であろう。声のオーバーラップやパラ言語的な部分の表記、言いよどみなど、会話には単純にテキスト化できないものが非常に多い。これをどう表すかが問題になる。本コーパスでは BTSJ を採用しているが、表記の方法と記号の体系については他にも様々なものがあり、研究の目的に応じて利用されているのが実情である。

各種コーディング（タグ付け）については信頼性の確保が重要である。コーディングが研究者個人ベースで行われると、心理的なバイアスや、何らかの恒常的な誤差などが生じる可能性があり、少なくともデータの一部を別の研究者にもコーディングしてもらい Cohen's κ などの一致度を算出して、信頼性を確認した上で個人による作業を行う必要がある。今回公開したデータでは「発話文」の認定についてのみコーディングの一致度を測定したが、それだけでも相当の労力が必要である。しかし、避けて通れないものだと思われる。

また、今回はインタビュイーの同意をもとに、収録データが公開できたが、それは事前にインフォームド・コンセントが得られるインタビューという場面だからであろう。完全に自然な会話場面の「かくし撮り」的な状況だと、当然ながら、個人情報の保護や倫理的な問題での点検が必須になる（参照：西郡, 2002）。

現在、研究組織による大規模データの収集の事例も見られ、その利用も推奨できるものであるが、個別的な目的を持ったデータ収集となると個人または研究プロジェクト単位で行わざるを得ない。その際には上記のような困難な部分があることを、指摘しかつ自覚したい。

【参考文献】

磯野英治（2009a）「日本語母語話者の会話におけるターン交代の定量的・定性的分析 ―インタビュー会話における調査から―」首都大学東京 人文科学研

究科 修士論文
磯野英治 (2009b)「日本語母語話者の会話におけるターン交替の特徴に関する定量的分析－インタビュー会話における調査から－」、口頭発表、2009 年度日本語教育国際研究大会、ニューサウスウェールズ大学、シドニー、オーストラリア
磯野英治 (2009c)「日本語母語話者のターン交替における定量的分析とその語用論的特徴について－会話教育への示唆－」2009 韓国 日本学会 傘下学会連合学術大会（漢陽女子大学、ソウル、韓国）
宇佐美まゆみ(2003)「改訂版: 基本的な文字化の原則 (Basic Transcription System for Japanese: BTSJ)」『多文化共生社会における異文化コミュニケーション教育のための基礎的研究』平成 13-14 年度科学研究費補助金基盤研究Ｃ(2) (課題番号: 13680351)(研究代表者: 宇佐美まゆみ) 研究成果報告書: pp.4-21
鎌田修(2006)「KY コーパスと日本語教育研究」日本語教育，130 号，pp.24-51
国際交流基金・日本国際教育協会(1994)『日本語能力試験　出題基準』
崔文姫 (2009a)「日本語学習者に対する日本語母語話者の印象形成」首都大学東京 人文科学研究科 博士論文
崔文姫 (2009b)「「好ましさ」の印象形成要因－ケーススタディを通して－」首都大学東京・東京都立大学 日本語・日本語教育研究会『日本語研究』29 号，pp.21-35
西郡 仁朗 (2002)「自然会話データ『偶然の初対面の会話』～その方法論について～」『人文学報』３３０号，東京都立大学人文学部，pp.1-18