International Corpus of Japanese as a Second Language

国語研プロジェクトレビュー
NINJAL Project Review　Vol.6 No.3 pp.93–110（March 2016）

多言語母語の日本語学習者横断コーパス

International Corpus of Japanese as a Second Language

迫田 久美子（SAKODA Kumiko），小西 円（KONISHI Madoka），
佐々木 藍子（SASAKI Aiko），須賀 和香子（SUGA Wakako），細井 陽子（HOSOI Yoko）

1. はじめに

1.1 プロジェクトの背景

本稿の目的は，現在，構築を進めている多言語母語の日本語学習者横断コーパス（I-JAS: International Corpus of Japanese As a Second Language）のデータ収集調査から書き起こしまでの概要を示すものである。

現在，海外には 398 万人の日本語学習者がいる（2012 年度日本語教育機関調査 国際交流基金）。また，国内に在住する外国人の数は，212 万人であり（平成 26 年度末 法務省），労働力確保，高度人財の登用も含め，日本は確実に多文化共生の社会に向かっている。共同研究プロジェクト「多文化共生社会における日本語教育研究」では，第二言語習得研究，対照言語学，社会言語学，心理言語学，コーパス言語学などの幅広い学問領域の連携により，第二言語としての日本語の教育・学習をめぐるさまざまな問題について，実証的な研究を行うことを目的としている。

共同研究プロジェクト「多文化共生社会における日本語教育研究」では，2012 年よりサブ・プロジェクトとして，海外で学ぶ日本語学習者のコーパスを構築するための調査とそれに伴う研究を開始した。本稿で使用する「プロジェクト」という用語は，この日本語学習者コーパス（I-JAS）構築のプロジェクトを指す。また，コーパスは「コンピュータで処理できるデータベース化された大規模な言語資料」と定義する。

言語研究の領域では，1990 年代に入り，コーパス言語学という新しい学問領域が誕生し，コーパスに基づく研究が盛んに行われるようになった。英語教育では，S. Granger の ICLE（International Corpus of Learner English）[1]，投野由紀夫の JEFLL（Japanese EFL Learner）[2]，石川慎一郎の ICNALE（International Corpus Network of Asian Learners of English）[3] など，すでにさまざまな英語学習者のコーパスが登場し，研究が進められている。

学習者のコーパスは，学習者の言語習得のみならず，教授法や言語研究に重要な意味を持つ。例えば，教師にとって，コーパスに見られる誤用は学習者の学習困難点を知り，その原

[1] http://www.uclouvain.be/en-cecl-icle.html
[2] http://jefll.corpuscobo.net/
[3] http://language.sakura.ne.jp/icnale/

因が母語の影響か，訓練上の転移か，共通のメカニズムによるものかを推測する手がかりとなる。データを見ると「先生，よろしくね」など，文法的には誤用とは言えないが，社会言語学的な待遇表現の観点から見ると不自然な表現が出現し，指導の工夫が必要であることが分かる。また，発達の指標にもなる。「のだ」や「受け身」などは，初級レベル後半で学習することが一般的であり，それらの誤用は初級レベル前半では観察されない。したがって，これらが出現することで習得が進んでいることが推測できる。

研究者にとっても，学習者の生のデータは研究の上で重要な材料となる。これまでの習得研究では，穴埋めテストや多肢選択，文完成法など規範的な目標言語を基本として調査が行われていたが，このような方法では学習者自身の文法は解明できない。学習者自身がどのような独自の工夫，ストラテジーを用いて文法を形成しているのかについての解明は，実際の彼らの言語使用を観察することから始まる。また，母語の影響を見るのであれば，1 ヵ国の母語の学習者のデータを見ただけでは主張できない。必ず，その言語体系とは異なる言語体系の母語 1 ヵ国以上の学習者のデータを比較分析することが必要となる。

さらに，コーパスのような大量のデータを扱うことで，日本語能力レベルと言語発達の関係，習得困難点の要因や言語習得のメカニズムの解明の糸口を見つけることができるのである。学習者コーパスは，そのような指導のヒントや研究の種が貯蔵されている貴重な宝庫だと言える。

1.2　I-JAS の誕生

共同研究プロジェクト「多文化共生社会における日本語教育研究」は，2013 年 1 月に中国語・韓国語母語の日本語学習者縦断発話コーパス（C-JAS: Corpus of Japanese As a Second Language）を公開した。中国語母語話者 3 名，韓国語母語話者 3 名の約 3 年間の縦断調査による発話データに形態論情報，誤用タグを付けた 46.5 時間分（約 57 万語）のコーパスである。当時，公開されているデータで特定の学習者を 3 年間追跡して会話データを収集し，形態素や誤用タグを付けたコーパスはなく，貴重なデータとして活用されている。

しかし，C-JAS にはいくつかの問題点がある。1 つは，学習者個人の日本語能力レベルの測定が行われていないために，学習者間の習得過程の比較ができない点である。もう 1 つは，母語が中国語と韓国語の 2 つのみであるという点である。第二言語習得研究で最も注目される要因は母語の影響である。C-JAS では 2 ヵ国，KY コーパス（表 1 参照）では英中韓の 3 ヵ国，他の発話コーパスを見ても，英中韓に東南アジアの数ヵ国が加わる程度で母語の異なる学習者のデータは極めて少ない。

これらの問題点を踏まえ，本プロジェクトは日本語学習者を対象に大規模な学習者コーパスを構築することを目的としてスタートした。構築予定のコーパス名称は「多言語母語の日本語学習者横断コーパス」（I-JAS: International Corpus of Japanese As a Second Language）である。

このコーパスは，約 1000 人の日本語学習者のデータを収集して構築する予定である。海外では，12 言語の異なる母語の学習者を対象とし，国内においては教室環境と自然環境の学習者および日本語母語話者を対象としてデータ収集を行った。コーパスは，書き起こしデー

タと付属の検索システム，音声データ，さらに一部，作文データを公開予定である。2016年春に225名分の第一次データを公開し，毎年度末に順次データを公開していく予定で，全てのデータの公開は2020年春を目標としている。

2. I-JASの概要

2.1 これまでの日本語学習者コーパス

I-JASの概要を述べるにあたり，まず，既存の日本語学習者コーパスを概観する。I-JASには日本語学習者の発話と作文のデータが収集されているが，発話の比重がより大きい。そこで，これまでの発話コーパスの内容を大まかに押さえておくこととする。それらを示したものが表1である。

表1 主な日本語学習者発話コーパス[4]

コーパスの名称	データ量（時間）	学習者の母語	縦断/横断	背景調査	データ収集方法	レベル判定	検索システムの有無
KYコーパス	90本（30分/1本）	中国語（30名），韓国語(30名),英語(30名)	横断	×	OPI	OPI	○
会話DB（横断編）	339本（30分/1本）	韓国語，中国語，英語，インドネシア語，その他	横断	○	OPI	OPI	×
会話DB（縦断編）	12〜25本（約30分/1本）	タガログ語，韓国語，中国語，ロシア語，マレー語，ポルトガル語	縦断（1〜5年）	×	OPI	OPI	×
BTSJ	57会話（約17時間）	韓国語，中国語，フランス語，その他	横断	×	対話： ・雑談 ・論文指導 ・電話	×	×
上村コーパス	66本（20〜30分/1本）	英語（27名），韓国語(22名),中国語(4名)，デンマーク語，ロシア語（各2名），その他9ヵ国（9名）	横断	×	OPI準拠： ・会話 ・ロールプレイ	×	○
発話対照DB	190名	中国語（69名），韓国語（70名），タイ語（51名）	横断	○	・朗読3課題 ・スピーチ4課題 ・ロールプレイ4課題	SPOT（一部）	×
LARP	37本（20分/月1×3年半）	中国語（37名）	縦断（3年半）	○	作文→フォローアップインタビュー	SPOT	×
C-JAS	47本（60分/1本）	中国語（3名），韓国語（3名）	縦断（3年半）	△	NSとの自由会話（テーマ有）	×	○

[4] 正式名称は稿末の参照コーパス一覧に示す。

データ量は，最大のもので 339 名分，各 30 分であり，学習者の母語の種類は中国語と韓国語が主になっている。学習者の属性，言語環境，日本語学習などに関する背景調査は行っていないものもある。背景調査を行っている場合でも，言語形成期の言語環境などの情報が不足しており，データから解釈が困難な場合があると思われる。データの収集方法は，OPI（Oral Proficiency Interview）を採用したものが半数に上っている。OPI 以外では，複数のタスクを行っているものがあるが，そのバリエーションは多くても 3 種類にとどまっている。日本語能力のレベル判定はないものもあるが，行われている場合でも 1 種類のテストによるものである。レベル判定は研究内容と密接に関わるため，1 種類の判定結果だけではコーパスを汎用的に使用することが困難な場合も少なくない。多様な研究ニーズに応えるためには，複数の選択肢がある方が望ましい。また，検索システムは備えていないものもある。なお，表 1 では日本語母語話者のデータ量については除外して示した。既存のコーパスの課題を以下にまとめる。

a）母語別・学習環境別のデータ数が少なく，データに言語の偏りがある
b）学習者の背景情報が不足している，または詳細ではない
c）レベル判定が，日本語能力の客観テストとしては不十分
d）タスクのバリエーションが乏しい
e）同一の学習者による発話と作文を備えたデータが少ない
f）検索システムを備えたものが少ない

I-JAS の構築においては，これら既存のコーパスの課題を踏まえ，1000 人を超えるデータ数，12 の異なる母語の学習者を対象とし，統一的な日本語能力テストを実施するなどのコーパス設計を行った。

2.2　I-JAS の特徴

2.2.1　調査対象者の人数と内訳

本コーパス構築のために行った調査の対象者について説明する。対象者は環境別に 3 種類に分類されるが，共通の条件として日本語を第二言語とする者で，日本語母語話者との 30 分の対話に対応でき，かつある程度読み書きが可能な学習者とした。対象者は，①海外の教室環境学習者，②国内の教室環境学習者，③国内の自然環境学習者である。①は海外 17 の国と地域，計 20 ヵ所の JFL（外国語としての日本語教育）環境の教育機関において，体系的に日本語を学んでいる学習者である。調査地は，インドネシア，スペイン，タイ，トルコ，ハンガリー，フランス，ベトナム，ロシア，英語圏（アメリカ，イギリス，オーストラリア，ニュージーランド），ドイツ語圏（ドイツ，オーストリア），韓国（2 ヵ所），中国語圏（本土 2 ヵ所，台湾 2 ヵ所）の計 20 ヵ所で，母語の種類は 12 言語である。この 12 言語は母語の影響についても分析できるよう言語類型論における分類を参考に決定した。②は初級の段階で来日し，日本における教育機関で 1 年以上日本語を体系的に学んでいる学習者で，主に

日本語学校の在学生や，日本の大学に留学している大学生である。③は日本で生活をしながら，自然に日本語を習得している学習者である。基本的には，体系的に日本語学習をしていないことを条件としたが，週 3 日以内のボランティア教室での日本語学習は許容とした。②③は，国内 8 ヵ所（東京都，静岡県，広島県）で調査を行った。

さらに，上記の日本語学習者に加え，20 代から 50 代の日本語教育経験のない日本語母語話者にも学習者と同様の調査を行った。

データ公開の予定人数および各言語類型は表 2，3 の通りである。

表 2　調査地および調査環境別の公開予定人数と言語類型[5]

	環境/母語		言語類型	公開予定人数（人）
1	海外（教室環境）	インドネシア語	オーストロネシア語族	50
2		スペイン語	印欧語族―イタリック語派	50
3		タイ語	カム・タイ語族	50
4		トルコ語	アルタイ語族	50
5		ドイツ語	印欧語族―ゲルマン語派	50
6		ハンガリー語	ウラル語族	50
7		フランス語	印欧語族―イタリック語派	50
8		ベトナム語	モン・クメール語族	50
9		ロシア語	印欧語族―スラブ語派	50
10		英語	印欧語族―ゲルマン語派	100
11		韓国語	不明	100
12		中国語	シナ・チベット語族	200
13	国内	教室環境	―	100
14		自然環境	―	50

表 3　日本語母語話者の年代別公開予定人数

20 代	30 代	40 代・50 代	合計
20 人	14 人	16 人	50 人

2.2.2　詳細な学習者情報

本調査では国内外の多様な日本語学習者のデータを収集するため，様々な学習環境や背景情報を詳細まで正確に収集する必要があった。収集方法はアンケート形式で，学習者の属性について 7 項目，学習者の現在の言語環境について 7 項目，日本語学習開始から現在までの日本語との関わりについて 6 項目の全 20 項目となっている。学習者の負担とプライバシーを配慮し，「答えたくない/答えられない」という選択肢も設けた。また，ウェブで入力可能

[5] 言語類型は角田（2009）を参照した。

な回答システムを作成し，対面調査の前に学習者に回答を依頼した。この学習者情報は情報を正確に収集するため，学習者の母語等で回答できるよう，全 10 種類（インドネシア語，英語，韓国語，スペイン語，タイ語，中国語：簡体字版，中国語：繁体字版，ハンガリー語，フランス語，ベトナム語）の翻訳版が作成された。回答システムでの入力漏れや不備があった際には，対面調査で直接本人に確認を行った。

2.2.3　日本語能力の客観テスト

調査実施時に学習者がどの程度日本語の言語知識を持っているかという全体像は不明であった。本コーパスでは，コーパスを利用する研究者それぞれの研究内容に応じてレベル判定基準を選択できるよう 2 種類の日本語能力の客観テストを行った。テストには J-CAT と SPOT を使用した。J-CAT（Japanese Computerized Adaptive Test）は日本語能力自動判定テストで，聴解，語彙，文法，読解の 4 セクションから日本語能力を測定するものである。SPOT（Simple Performance-Oriented Test）は TTBJ（Tsukuba Test-Battery of Japanese）の 1 つで，言語知識と言語運用の両面から日本語能力を測定するものである。前者は，1 時間程度，後者は 15 分程度で行うことができるものである。これらのテストはウェブ上で公開されており，誰でも自由に受験することが可能であるが，海外でのインターネット環境の安定しない調査地に備え，テスト作成者の支援協力を得て，インターネットを使用せず受験できるよう，PC にシステムを搭載して調査地で実施した。

2.2.4　タスクのバリエーション

本コーパスは，1 人の学習者に対して 7 種類 12 のタスクを行っていることも大きな特徴の 1 つである。タスクは，①ストーリーテリング（2 タスク），②対話（約 30 分間），③ロールプレイ（2 タスク），④絵描写，⑤ストーリーライティング（2 タスク），⑥メール文（3 タスク），⑦エッセイ，の 12 タスクである。このうち，①から⑤は対面調査で調査実施者である日本語母語話者が対応し，⑥と⑦は任意で対面調査の事前に行った。

調査は，調査地ごとに複数名の協力者で担当したため，できる限りどの調査地でも同条件となるよう，調査マニュアルを作成した。そして，対話の進め方など対面調査上の注意点について，共通認識を持った上で調査が実施できるよう調査実施者に対して調査に関する事前研修も実施した。

2.3　調査の内容のバリエーション

2.3.1　対面調査のタスク詳細

調査は，以下の順序で行われた。それぞれのタスクについて，詳細を説明する。

1）ストーリーテリング　提示されたイラストのストーリーを話すというもので，図 1，2 の 2 種類のイラストを用いて行った。登場人物の名前や日本語能力試験の旧 2 級以上の名詞語彙には，日本語と英語の訳を付与した。タスクを始める前には，ストーリーの内容を確認する時間を 1 分以内で設けた。このタスクは，対話で十分な発話量を引き出せない場合への対

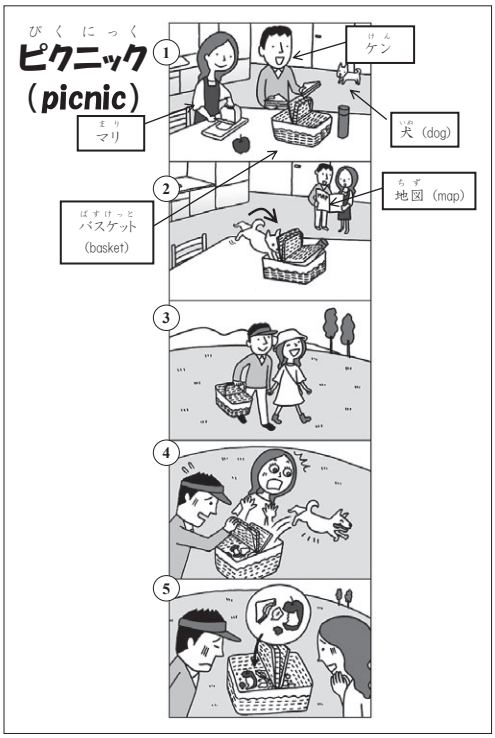

図1　ストーリーテリング　イラスト1

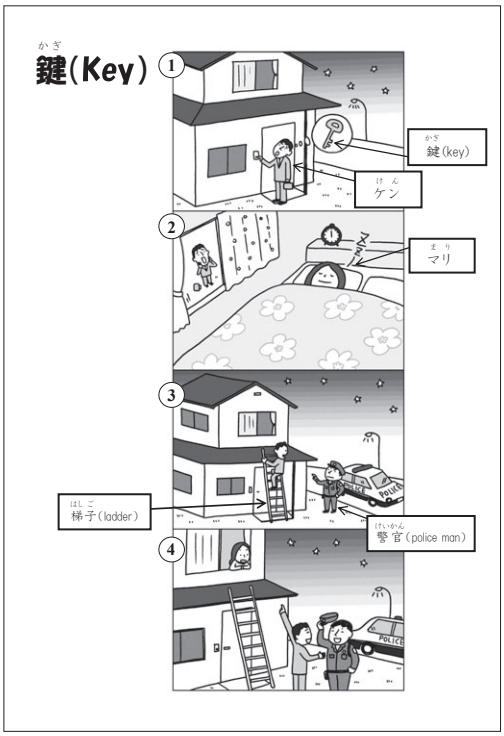

図2　ストーリーテリング　イラスト2

策，また，談話レベルの発話ではあまり観察できない文法項目（受身や自他動詞，複合動詞など）の使用状況を観察することを目的として実施した。日本語能力レベルの低い学習者でもイラストを手掛かりに発話することができ，ある程度まとまった談話を引き出すことが可能である。また，このタスクを対面調査の最初に行うことで，学習者の緊張を和らげる効果もあった。イラストは先行研究で使用されているものを参考にしたが，海外で使用することも配慮し，イラストレーターに作成を依頼した。

2）対話　学習者と調査実施者が自然な会話を 30 分程度行うものである。データ間での比較を考慮し，内容はある程度統一された話題を設定し，自然な会話の流れで学習者の言語運用を引き出すようにした。自然な流れで会話を進めるという手法は OPI を参考にしたが，本調査では日本語の客観テスト以外では評価を行わないため，対面調査において評価は行っていない。話題は全 15 項目で，前半は学習者本人に関する過去，現在，未来について話せるものをバランスよく配置し，後半には，談話レベルの発話が引き出せるよう，意見陳述や反論ができるような話題を設定した。

3）ロールプレイ　設定された場面に応じて，与えられた役を演じて会話するタスクであるが，本調査では日本料理店でのアルバイト場面を設定し，店長にアルバイトの出勤日数の変更を依頼するものと，店長からの仕事内容の変更依頼を断るものの 2 種類を実施した。このタスクは，学習者の日本語によるコミュニケーション能力，交渉能力を観察する目的で実施

された。タスクの検討に際しては，初級から上級レベルの幅広い学習者に対して実施可能なこと，学習環境の異なる国内外の学習者にも対応可能なことを考慮した。また，ロールカードは，内容を正確に理解できるよう学習者情報の回答システムと同様の 9 言語・10 種類に翻訳したものにポルトガル語の資料を追加し，全 11 種類の翻訳資料を使用した。

図 3　絵描写 イラスト

4）絵描写　図 3 のイラストを見て，説明するものである。イラストは，許（1997）の研究で使用されたものを著者の許可を得て使用した。日本語能力レベルの低い学習者にも達成感を与えられるよう配慮し，家の中のこと，家の外のことに分けて口頭産出を促した。このタスクは，対話では見られない動詞の活用に着目する目的で行われたが，調査実施期間の途中から追加されたタスクであったため，一部実施していない調査地もある。

5）ストーリーライティング　ストーリーテリングと同一の 2 種類のタスクを PC で入力するというものである。対面調査の最後に行い，最初に行ったストーリーテリングから，40〜50 分後に実施した。日本語での PC 入力が困難な場合は手書きで対応した。制限時間は 1 タスクにつきおよそ 10 分で，辞書やインターネットの使用は不可とした。このタスクは同種のタスクを口頭産出した場合と文産出した場合の相違点を探る目的で実施した。

2.3.2　作文調査（メール文・エッセイ）の詳細

作文調査は対面調査に協力してくれた学習者の中から，任意で行った。この調査は同一の学習者による，発話と作文を比較することを目的としている。対面調査前にタスクを与え，学習者は各自自宅や学校などで行った。タスクは実際に学習者が日本語で書く可能性がある状況を想定したもので，メール文 3 タスク，エッセイ 1 タスクであった。辞書，インターネットの使用は可能で，時間制限も設けていないが，日本人や日本語教師に尋ねたり，助けを求めたりしないよう指示した。このタスクのインストラクションも条件が同じとなるよう，翻訳版全 11 種類（ロールカードと同様の翻訳言語）を使用して実施した。メール文のタスクは①教師に奨学金のための推薦状を依頼する，②教師に期日までにレポートを提出できないことを伝える，③教師に自国の案内を依頼されたが，用があるため対応できないことを伝えるというもので，エッセイは「私たちの食生活：ファーストフードと家庭料理」というタイトルで 600 字程度のものであった。

また，作文調査実施後にはアンケートも実施した。アンケートの内容は参照資料の有無，タスクに要した時間，母語や日本語における文章構成に関するこれまでの学習の有無などである。

同一の学習者の発話データと作文データを備えたコーパスはこれまであまり見られないため，I-JAS の大きな特徴の 1 つと言える。

2.4 調査の倫理

本プロジェクトの調査を実施するにあたり，国立国語研究所，調査実施機関および大学での倫理規定に基づき申請を行い，承認を得た。また，調査に参加する日本語学習者には調査概要を説明し，本プロジェクトと学習者双方に同意書を残す形で了承を得た。

3. 書き起こしとタグ付けの概要

3.1 書き起こしの基本的なルール

発話データを検索システムに搭載するためには，書き起こしを行う必要がある。発話の書き起こしやタグ付けは，その目的に応じて多様なルールがあり，コーパスのための書き起こしやタグ付けに限ってみても，小磯他（2006），迫田他（2014）などがある。しかし，異なる目的を持つ I-JAS に対して，それらをそのまま用いることは妥当ではない。そのため，I-JAS の目的に沿った書き起こし方針を定める必要がある。

I-JAS の書き起こしの方針としては，①日本語学習者の文法習得，談話習得などの研究を目的とした書き起こしであること，②学習者，日本語母語話者の両方の発話をできるだけ発音に忠実に書き起こすが，発話の重なりやポーズの長さ，声の大きさなどの情報は付与しないこと，という 2 点があった。書き起こしにおける基本的なルールは以下の通りである。

1）表記　表記は特段複雑なルールは設けず，一般的な漢字仮名交じり文で行った[6]。特別なルールとして定めたのは，①数字は漢数字で記す，②擬音語・擬態語はカタカナで記す，③外国語はどの言語であってもカタカナを基本とする[7]（例：「クライムする」「グアンジー省」），④アルファベット通りの発音をしている場合はアルファベット全角で記す（例：「ディーエス」と言っている場合は「DS」と記すが，「ポップ」と言っている場合は「POP」とは書かずカタカナで書く），⑤漢字表記に複数の読みがある場合は「一日（いちにち）」「一日（ついたち）」のように括弧付きで振り仮名を付す，というものである。

2）文の区切り　発話音声で文の区切りを定めるのは大変難しい作業である。本コーパスでは，文区切りの位置で改行することとし，①対話形式のタスク（対話とロールプレイ）は話者交代で改行，②独話形式のタスク（ストーリーテリングと絵描写）は 1 文と思われる単位で改行し，タスク中に対話部分があればその部分は話者交代で改行とした。話者交代の位置を文区切りの位置とみなす箇所がある関係上，1 文の終わりを示す「。」は使用せず，ポーズがある箇所に「、」を打つにとどめた。ポーズや伸ばす音の秒数はカウントしていない。上昇イントネーションがあると判断した箇所には「？」を付けた。したがって「？」は文末とは限らない。

---

[6] 表記についてルールを緩やかにしたのは，「上る」「登る」「昇る」など異表記であっても形態素解析結果には同じ語彙素が付されるためである。しかし，タスクによっては一部表記を固定した箇所もある。

[7] 中国語のみ，日本で通常漢字で書くと思われる語は漢字表記とした。「上海」「北京」などである。

3）あいづちと発話の重なり　あいづちとみなされる発話は，〈　〉で相手の発話の中に挿入する。発話が重なっている場合は，別の発話として改行して表記するか，あいづちとして相手の発話中に挿入する。〈　〉の中の表現は，形態素解析にかけた際，全てまとめて「あいづち」という品詞情報が付くようにした。

4）その他　音声が聞きとれない場合はおおよその音素数を「＊」の数で示した。映像を収集していない本コーパスでは非言語行動を特定することは難しいが，笑いや咳が 1 つのターンを取っている場合やそれらの非言語行動が会話に影響している場合もあるため，分かる範囲で｛　｝を用いて「｛咳｝」のように記載した。また，個人情報は「【人名 1 】さん」のように伏字化し，音声データにおいてもマスキング処理を行った。形態素解析にかけた際，非言語行動の品詞情報は「非言語行動」，個人情報の品詞情報は「個人情報」となる。

3.2　形態素解析のためのタグ付け

3.2.1　学習者発話の形態素解析の課題と対策

I-JAS の特徴の 1 つとして，文字列検索だけでなく，形態論情報を基にした検索も可能なかたちで公開される，という点がある。形態論情報を基にした検索を実現するためには，書き起こししたデータを形態素解析する必要がある。例えば「きのうはずっと勉強しました」を解析すると表 4 のようになる。自動の形態素解析器は，書き言葉であればかなり高い精度で解析ができるが，発話音声には言い淀み，あいづち，言い間違い，倒置など，多くの音声特有の現象があるため，書き言葉に比べるとやや精度が落ちる（小木曽 2014）。学習者の発話にはさらに発音や活用の誤り，予測不能な誤用，意味不明の語，多様な外国語，母語話者よりも多様なフィラーや語の断片が現れるため，それをそのまま解析すると誤解析が起こる。表 5 は「きのは、ぞっと、えと、べ、勉強しました」という発話の下線部に対する解析結果である。学習者の発話通りに書き起こしをして解析した場合，学習者の意図と異なる解析結果（表の網掛け部分）になってしまう。そのため，何らかの対策を取らねばらないことが分かる。

表 4　解析結果（一部抜粋）

表層形	語彙素	品詞大分類	品詞中分類
きのう	昨日	名詞	普通名詞
は	は	助詞	係助詞
ずっと	ずっと	副詞	
勉強	勉強	名詞	普通名詞
し	為る	動詞	非自立可能
まし	ます	助動詞	
た	た	助動詞	

表 5　学習者発話の解析（一部抜粋）

表層形	語彙素	品詞	学習者の意図
き	来る	動詞	昨日
の	の	助詞	
は	は	助詞	
、	、	補助記号	
ぞっと	ぞっと	副詞	ずっと
、	、	補助記号	
えと	干支	名詞	フィラー
、	、	補助記号	
べ	べい	助詞	「勉強」の断片

I-JAS に求められる研究のニーズには，2 つの方向性がある。1 つは，学習者のありのままの発話を研究対象とするものであり，その場合に必要なのは，発話を忠実に書き起こしたデータである。そのため，表 5 の表層形の部分のみを抽出したデータ「発話プレーンテキスト」を作成し，公開する。もう 1 つのニーズは，学習者の意図を反映した形態論情報が付与されたデータを検索・収集して研究対象とするものであり，その場合に必要なのは，誤解析を誘発する部分に対して学習者の意図を汲んだ形態論情報を付与した解析結果である。そのような結果を得る方法の 1 つとして，表層形に対して学習者の意図やそれを表すタグを人手で付与していく方法がある。しかし，誤解析を誘発する部分の同定は人によって差があり，また，学習者の発話には複数の解釈がありうる場合や解釈不能の場合がある。全てのデータに対して同じ基準で判断を行うことは容易ではないため，明確なルールが必要である。そのため，I-JAS では独自のタグを設計した。その際，「誤解析を誘発する部分に対して，学習者の意図を汲んだ形態論情報を付与できる方法」，「人手で作業を行うことが可能な範囲で十分な効果が出る方法」という 2 つの方針を設けた。「タグ付きテキスト」は，「発話プレーンテキスト」にタグを付与するかたちで作成した。

書き起こしやタグ付けは，学習者の発話を理解しつつ行う必要があるため，日本語教育経験のある者や日本語教育を学んだことのある学生・院生をアノテーター[8] とし，全て人手で行った。全てのデータに対して統一した基準で書き起こしやタグ付けを行う必要があるため，複数回の研修を実施した。また，書き起こし，タグ付けされたデータに対して必ずもう 1 名のアノテーターがチェックする体制をとり，タグ付けのミスの起こりやすい箇所に対してさらに研究員が 2 回のチェックを行うこととした。

3.2.2 I-JAS のタグ

表 6 に示す通り，発話データに付与するタグは全部で 8 種類ある。その処理のタイプは，①解析できない箇所を解析から除外する，②誤りを含むために誤解析になりやすい箇所に解

表 6 I-JAS のタグ

処理	内容	タグ表記
解析から除外	意味不明語 語の断片	[α＝X]
解析用の語を指定	語中の長音，ポーズ	[α＝T＝β]
	語や活用や発音の誤り	[α＝G＝β]
解析用の品詞を指定	フィラーを感動詞に指定	[α＝F]
	外国語を名詞に指定	[α＝N]
	連体詞に指定	[α＝R]
曖昧性への対応	発音不明瞭（α1 か α2）	[α1/α2＝H]
	複数の読みがある漢字語 A	[α（読み）＝Y]

[8] ここでは，書き起こし・タグ付けの作業者を「アノテーター」と呼ぶ。

析用の語を指定し，それを解析対象とする，③解析で間違った品詞が付きやすい箇所に解析用の品詞を指定し，それを解析対象とする，④曖昧性への対応を行う，の 4 つに分けられる[9]。

以下に，タグを付与した発話の例を示す。

（1）［きの＝G＝昨日］は、［ぞっと＝G＝ずっと］、［えと＝F］、［ベ＝X］、勉強しました
（2）［いちばーん＝T＝一番］、最初、［へ＝X］、部屋、［も / もう＝H］［開いて（あいて）＝Y］
（3）［パッキャラマコ＝N］というー、［んー＝F］、［その＝R］［むすそー＝G＝仏像］

（1）では，「きの」という発話に「語や活用や発音の誤りに対して，解析用の正しい語を指定する」働きを持つタグ G が付され，「昨日」という修正が行われている。形態素解析は，修正した「昨日」に対して行われるため，「きの」という発話に「昨日」という語彙素（見出し語）が付与されることになる。「ベ」は「勉強」という語の言いかけ（語の断片）であると考えられるが，そのようには解析されず誤解析となるため，解析から除外する。タグ X が付与された語の品詞情報は「解析困難箇所」となる。また，学習者の多様なフィラーは誤解析を誘発するため，あらかじめタグ F を付与しておく。品詞は感動詞となる。音声的に曖昧な箇所は（2）のように「も / もう」と複数列挙してタグ H を振り，最初に記載した「も」に対して形態素解析を行う。また，「開いて」のように「あいて」「ひらいて」の複数の読みが生じる漢字語に対しては，タグ Y を付与した上で（　）を用いて振り仮名を振る。この場合，（　）の中の文字は解析対象から除外されている。（3）の「パッキャラマコ」はタイの仏像の名前であるが，形態素解析辞書（UniDic）に登録されていない外国語のような語は誤解析になる可能性があるため，タグ N を付与する。タグ N を付与した語は，それが形態素解析辞書に登録されている語であればその形態論情報が付くが，辞書にない場合は 1 語の未知語と判断され，品詞は名詞となる。また，外国の地名や料理名などの詳細をアノテーターが把握できた場合は，「タイの仏像の名前」のような補足情報を付与しておき，検索時に表示されるようになっている。また，「この」「その」「あの」の連体詞は，感動詞と誤解析される場合もあることから，発話を聞いて連体詞と判断できた場合にはタグ R を付与し，品詞を連体詞と固定することとした。

このようなタグを付与した箇所は，形態論情報を基にした検索を行う際に，より学習者の意図に近い検索を行うことが可能になる。例えば，「ずっと」の検索結果として（1）も含まれ，「番」の検索結果として（2）も含まれる。また，タグによる検索も可能であるため，タグ G を検索するとアノテーターに語や活用や発音の誤りと判断された箇所が検出される。同様に，タグ X を検索すると語の断片や意味不明語と判断された箇所が検出される。これらの検索結果には，ある特定のタイプの誤用や，学習者の言い直しに関する情報がまとまっていることになる。

[9] I-JAS の解析精度については小西他（2015）に詳しい。本論で述べたタグで十分な解析精度を維持している。

3.2.3 注意すべきタグ付けの例

タグを付与する過程で大きな問題となるのは，文脈からも学習者の意図が正確に把握できず，修正候補が 1 つに決められない場合である。例えば，「宿題がもでません」という発話は，文脈からは「宿題が持てません」「宿題がもう出ません」「宿題を持っていません」などの修正候補が想定されたが，1 つに絞ることはできなかった。このような場合は，1 つの候補に決定することを避け，「もでません」の箇所を「解析困難箇所」とみなしタグ X を付与することとした[10]。しかし，「もでません」に上記のような候補があることをユーザーに示すことは有効だと考えたため，修正候補は補足情報として検索時に表示されるようにしている。

また，学習者は，正しい語を発話しようとして何度も同じ語を言い直したり，正しい活用を思い出すために多様な語の断片を発話したりする。その場合，語がまとまった形で最後まで発話されていれば，タグを付与せずそのまま解析をするが，語が断片として現れている場合や，語や活用や発音の誤りが含まれている場合はタグによる修正をすることとした。

(4) 食べません、食べません、食べませんでした
(5) えっと、叱る、[叱れ＝X]？[叱れれた＝G＝叱られた]、叱られたんです
(6) サンドイッチの箱を、[飛び＝X]、[飛び込みます、みました＝G＝飛び込みました]

(4) は 3 語とも完全な形で発話されているが，(5) は「叱れ」は語の断片，「叱れれた」は発音（活用）誤りとして処理した。また，学習者の言い直しは日本語母語話者とは異なる位置で行われることもある。(6) のように，「飛び込みました」を発話しようとして「飛び込みます、みました」と発話している場合，このまま解析をすると「みました」は「見ました」になってしまう。そのため，まとまった語が部分と部分に分かれており，つなげればまとまった語になる場合は，つなげた形でタグ G による修正を行うこととした。しかし，「飛び」のように部分しか発話できていない場合は語の断片と判断し，タグ X となる。

(7) テーブルの上の、[あ＝F]、[みんなの＝G＝瓶など] 入れ物
(8) えーと、[おに＝X]？姉、姉？兄？兄かな姉かな、姉か
(9) [とーもだち＝T＝友達] とー [がっこ＝G 学校] が行きました

さらに，学習者の発音間違いの結果が，たまたま別の日本語の語になる場合もある。(7) は，テーブルの上に瓶が置かれている絵を描写している発話であるが，発話音声を聞くと「瓶など」が「みんなの」と聞こえる。このような場合は，調査で用いた絵から「瓶」であると判断し，「瓶など」に修正する。しかし，(8) のように「姉」「兄」という語を思い出すために口から出た「おに」という語は「鬼」とは考えにくいため，意味不明語と考える。

[10] このような例からも，学習者の意図を汲んだ修正がアノテーターの主観を含んだものであることが分かる。発話を書き起こすという段階においても主観を排除することはできない。そのような主観を極力統一したルールに近付けるために複数回のチェック体制をとったが，それでも揺れを完全に排除することはできない。この点は留意が必要である。

タグの設計の方針として，「誤解析を誘発する部分に対して，学習者の意図を汲んだ形態論情報を付与できる方法」という方針があるが，これは，学習者の誤用を全て修正するということではない。例えば，（9）の場合，「とーもだち」や「がっこ」は，学習者の発話意図が「友達」「学校」であると判断でき，かつ，その部分が誤解析になる可能性が高いため，タグによる修正を行うが，「学校が行きました」における「が」は修正しない。文法規則上正しい日本語を考えると，「が」は「に」か「へ」であるべきであるが，この発話では学習者は助詞「が」を言おうとして「が」と言っていると判断され，また，その通りに解析される。つまり，I-JAS のタグは，学習者の全ての誤用を修正するのではなく，学習者の意図と異なる形で解析されたり，正しく発話できていないがゆえに誤解析になると思われる箇所を修正することを基本としている。それは，学習者の誤用修正は研究者一人一人によって答えが異なるほど多様なものであり，その最終判断はユーザーにゆだねるべきと判断したためである。そのため，タグ G や X を検索することで，学習者の誤用が全て入手できるわけではないことには十分な注意が必要である。

なお，I-JAS 側で判断が必要だった箇所もある。「大きいなビーチ」「作りたくないのとき」「かわいいだから」のような過剰使用と判断される「な」「の」「だ」は，補足情報に「過剰使用」と付与した上で，「な」は助動詞「だ」，「の」は格助詞「の」，「だ」は助動詞「だ」という解析結果を指定した。これも，誤解析を避け，統一した基準で結果を示すためである。

4. おわりに

本稿は，「はじめに」で多言語母語の日本語学習者横断コーパス（I-JAS）がどのような経緯で立案され，日本語指導においても研究においても意義があることを述べた。次に，「I-JAS の概要」では，これまでの日本語学習者発話コーパスとは異なっている I-JAS の特徴を示し，具体的な内容について解説した。次に，「書き起こしとタグ付けの概要」では，対話やロールプレイ等の発話データの書き起こしやタグ付けの規則について解説した。これは，学習者言語や習得研究に興味がある人たちにとって，I-JAS を有効に利用するための必須情報である。

このコーパスは，日本語学習者 1000 人のデータ所蔵を目標としており，2016 年春には，第一次データ 225 人分の公開を目指している。2012 年の本プロジェクトのスタートから 4 年を経た現在，このプロジェクトの時間的経緯は 3 つの段階に分けられる。第一段階は，2012 年度の 1 年間を費やしてグランドプランを策定した時期である。対話調査者の調査方法の統一性を図るために，調査実施者への事前研修を実施したり，全てのタスクを何度も作り直し，できるだけ指示文を母語に訳したり，国内外でのパイロット調査も実施して準備に時間を費やした。

第二段階は，2013 年と 2014 年の 2 年間にわたり，20 ヵ所以上で海外・国内調査を実施した時期である。調査は学習者 1 人に対して 3 時間を要する。調査地 1 ヵ所につき 50〜60 人のデータ収集を，2 名の調査者が 5 日間で行った。調査地によっては，複雑で厳しい倫理申請の手続きを行い，学習者募集と調査遂行には現地の先生方に多大な労力をかけ，大変お世

話になった。

そして，第三段階は，2015 年度の書き起こしとタグ付けの段階である。関係者で何度も会議を重ね，国立国語研究所の他のコーパスを参照し共通理解を図りつつ，書き起こしやタグ付けのルールを作成した。アノテーターを日本語教育関係の院生や教師に限定して募集し，複数回の事前研修や練習期間を設けた。

このプロジェクトで構築される日本語学習者の発話コーパスは，2015 年現在，国内，海外を含め最も大きいコーパスである。2016 年春から順次，公開を目指していくが，我々のプロジェクトの最終目標は，コーパスの構築ではない。このコーパスを使って，どのような研究ができるのかを示していくことであり，多くの教師や研究者の方々がコーパス分析の面白さ，奥深さ，広がりを理解し，コーパスを利用した研究が発展することにある。

このプロジェクトをきっかけに，方言，日本語母語話者の現代語，歴史などのコーパスと連携を図り，それぞれの分野との研究の繋がりが検討されている。さらには中国語や英語を第二言語として学ぶ学習者コーパスのプロジェクトとも連携が築かれつつあり，コーパスに基づいた実証的研究の発展が十分に期待できる。本コーパスがその一翼を担うことができればと願う。

●付記●

本稿で示した多言語母語の日本語学習者横断コーパス（I-JAS）の開発には，非常に多くの日本語教師・日本語教育研究者の方々に支援をいただいた。国内外の調査に関わってくださった先生方，過酷な調査を遂行してくださった先生方，根気強く書き起こしやタグ付けを行っているアノテーターの皆さま，そして長時間にもかかわらず調査に参加してくださった 1000 人以上の学習者・日本語母語話者の皆さまには，深く感謝し，ここにお礼を申し上げます。

●参照文献●

バーナード・コムリー（著），松本克己・山本秀樹（訳）(1992)『言語普遍性と言語類型論：統語論と形態論』東京：ひつじ書房.

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

小西円・須賀和香子・佐々木藍子・細井陽子・八木豊・迫田久美子(2015)「学習者発話の高度な形態素解析を目指したタグの設計」『日本教育工学会第 31 回全国大会予稿集』943−944.

小木曽智信(2014)「形態素解析」山崎誠（編）『書き言葉コーパス 設計と構築』89−115．東京：朝倉書店.

迫田久美子・佐々木藍子・小西円・李在鎬(2014)『C-JAS（Corpus of Japanese as a second language）構築に関する報告書』国立国語研究所基幹型プロジェクト「多文化共生社会における日本語教育研究」報告書.

許夏珮(1997)「中・上級台湾人日本語学習者による『テイル』の習得に関する横断研究」『日本語教育』95: 37−48.

角田太作(2009)『世界の言語と日本語：言語類型論から見た日本語』東京：くろしお出版.

●参照コーパス●

・KY コーパス（鎌田修・山内博之）
　http://opi.jp/shiryo/ky_corp.html（説明），http://jhlee.sakura.ne.jp/kyc/（タグ付き）
・会話 DB（横断編）：日本語学習者会話データベース（国立国語研究所）
　https://nknet.ninjal.ac.jp/nknet/ndata/opi/
・会話 DB（縦断編）：日本語学習者会話データベース 縦断調査編（国立国語研究所）
　https://nknet.ninjal.ac.jp/judan_db/
・BTSJ：BTSJ（Basic Transcription System for Japanese）による日本語話し言葉コーパス（宇佐美まゆみ）
　宇佐美まゆみ監修（2011）『BTSJ による日本語話し言葉コーパス（トランスクリプト・音声）2011 年版』
　http://www.tufs.ac.jp/ts/personal/usamiken/btsj_corpus_explanation.htm（説明）
・上村コーパス：インタビュー形式による日本語会話データベース（上村隆一）CD-ROM
・発話対照 DB：日本語学習者による，日本語・母語対照データベース
　http://contr-db.ninjal.ac.jp/speech_01.html
・LARP: LARP at SCU（Language Acquisition Research Project at Soochow University）（台湾東呉大学）
　http://webbuilder.scu.edu.tw/builder/web_page.php?web＝156&pid＝9346
・C-JAS：中国語・韓国語母語の日本語学習者縦断発話コーパス（Corpus of Japanese As a Second language）（国立国語研究所）
　https://ninjal-sakoda.sakura.ne.jp/lsaj/

《要旨》本稿は，共同研究プロジェクト「多文化共生社会における日本語教育研究」が進めている多言語母語の日本語学習者の横断コーパス（通称 I-JAS）について概説した。

前半では，I-JAS 構築の経緯と概要，調査の内容と特徴をまとめ，後半では，I-JAS を利用する際に重要となる書き起こしのルールやタグ付けの方針などについて述べた。12 の異なる言語を母語とする約 1000 人の日本語学習者のコーパスは，日本語の第二言語習得研究や対照言語学，社会言語学的な言語研究のみならず，日本語教育の現場でも利用が期待される。

Abstract: This paper provides a description of I-JAS（International Corpus of Japanese As a Second Language）, which contains cross-sectional research data from Japanese language learners with different mother tongues. This corpus is a part of a collaborative research project entitled 'Study on Teaching and Learning Japanese as a Second Language in a Multicultural Society.'

The first half reports on the development of I-JAS and its salient features. The latter half describes the transcription rules and the basic principles of tagging, both of which are important for searching and extracting data from the corpus.

I-JAS includes data from approximately 1000 learners with 12 different native languages, and it will be a rich resource, not only for linguistic research in areas such as second language acquisition, contrastive analysis and sociolinguistics, but also for teaching Japanese as a second language.

迫田 久美子（さこだ・くみこ）

国立国語研究所日本語教育研究・情報センター教授。博士(教育学)(広島大学)。広島女学院中学・高等学校非常勤講師，広島大学教育学部非常勤講師，広島大学大学院教育学研究科教授を経て，2012 年 4 月より現職。
主な著書・論文：『日本語の中間言語研究―日本語学習者による指示詞コソアの習得研究―』（渓水社，1998），『日本語学習者の文法習得』（共著，大修館書店，2001），『日本語教育に生かす第二言語習得研究』（アルク，2002），『講座・日本語教育学 第三巻 言語学習の心理』（編著, スリーエーネットワーク, 2006），『日本語教育のためのコミュニケーション研究』（共著，くろしお出版，2012）など.
受賞：第 1 回日本語教育学会奨励賞（日本語教育学会，2003）.
社会活動：国際文化フォーラム理事，文化庁国語科審議会委員，国際交流基金諮問委員会委員，2015 年度外国人による日本語弁論大会審査委員長，2015 年度国際交流基金賞審査委員.

小西 円（こにし・まどか）

国立国語研究所日本語教育研究・情報センター プロジェクト非常勤研究員。博士（日本語教育学）（早稲田大学）。早稲田大学日本語教育研究センター助手などを経て，2011 年 7 月より現職。
主な著書・論文：「実態調査からみた「義務の表現」のバリエーションとその出現傾向」（『日本語教育』138, 2008），『日本語教育文法のための多様なアプローチ』（共著，ひつじ書房，2011）.

佐々木 藍子（ささき・あいこ）

国立国語研究所日本語教育研究・情報センター プロジェクト非常勤研究員。立教大学日本語教育センター兼任講師。修士（日本語教育学）（広島大学）。又松大学(大韓民国)専任講師，広島工業大学および日本語学校等の非常勤講師，また，2010 年 9 月から国立国語研究所日本語教育研究・情報センター プロジェクト奨励研究員を経て, 2014 年 3 月より現職。
主な著書・論文：「日本語学習者の接続助詞「から」の習得過程に関する研究―接続形式の習得に着目して」（『教育学研究紀要』52(2), 2006），「シャドーイング活動におけるリスニングの意義と役割―日本語短期集中コースでの実践報告」（共著，『広島大学日本語教育研究』21，2011）.

須賀 和香子（すが・わかこ）

元 国立国語研究所日本語教育研究・情報センター プロジェクト非常勤研究員。修士（日本語教育学）（早稲田大学）。ウドーンターニー・ラチャパット大学（タイ）専任講師，日本語学校および早稲田大学日本語教育研究センター等の非常勤講師を経て，2011 年 7 月より 2015 年 12 月まで在職。
主な著書・論文：『日本語教師のための「活動型」授業の手引き―内容中心・コミュニケーション活動のすすめ』（共著，スリーエーネットワーク，2008），『敬語使い方辞典』（共執筆者，新日本法規，2009）.

細井 陽子（ほそい・ようこ）

国立国語研究所日本語教育研究・情報センター プロジェクト非常勤研究員。修士（日本語教育学）（早稲田大学）。日本語学校常勤・非常勤講師を経て 2014 年 4 月より現職。
主な著書・論文：『みんなの日本語初級 I 漢字練習帳』第 2 版（東京国際日本語学院編著，スリーエーネットワーク，2012）.

基幹型共同研究プロジェクト「多文化共生社会における日本語教育研究」
プロジェクトリーダー　迫田久美子
（国立国語研究所 日本語教育研究・情報センター 教授）

プロジェクトの概要

本プロジェクトは，第二言語習得研究の枠組みを基盤としつつ，言語心理学，対照言語学等の関連諸領域との協働により，日本語学習者の言語環境と日本語の習得過程との関係を実証的に解明しようとするものである。具体的には，共同研究プロジェクト「多文化共生社会における日本語教育研究」は「日本語習得研究班」と「定住外国人の言語使用 / 複数言語使用と言語環境に関する研究班」から構成される。前者は，日本語非母語話者の日本語習得を自然環境か教室環境か，日本国内か海外か，などの環境要因および彼らの母語の違いが習得にどのように影響を与えるのかについて，学習者の会話コーパス・作文コーパスに基づいて明らかにする。後者は，多言語・多文化化が進む現代の地域社会における定住者や研究が進んでいない少数派の外国人の言語習得，複数の言語使用の実態をより的確に捉え，どのような日本語を必要とするのかを明らかにし，言語使用と言語生活の関係を明らかにする。