『言語情報学研究報告』No.10（2006）

日本語の学習者言語コーパスに関する基礎調査

古川　明子
（東京外国語大学大学院博士前期課程）

はじめに

本稿は「21 世紀 COE プログラム言語運用を基盤とする言語情報学拠点」の一環としての『上級学習者の日本語作文データベース』（海野多枝監修（2005））の公開・出版に向けて行われた基礎調査を記したものである。こうしたデータベース使用の研究の背景にあるコーパスの位置づけを確認し，特に日本語教育の立場から見た学習者言語コーパスの実態について調査を行った。

第二言語習得に関する研究のうち，実証的方法を用いた研究では学習者言語に関する調査およびデータの収集が必要とされる。しかしながら個々の研究の際に発話資料や作文，アンケート等の形でデータを集める場合，さまざまな条件からデータ自体が量あるいは質の面で不十分であったり，目的とする情報が得られなかったりすることもある。研究・分析の前段階としてデータ収集の方法や条件（調査対象・場所・時期等）についての綿密な計画と，その実行を支えうる条件が整うことが望ましいが，必ずしも調査研究に必要な要因が揃うとは限らない。

そうした環境を向上させる試みは昨今の IT 事情の発展と共に積極的に図られるようになった。言語研究に限ってみても，前提となるデータ収集を体系的に行い，当該の研究分析のみならず他の研究にも使用できるよう整備されたデータを作成するという手法が可能となった。さらには，データ収集自体を目的とし，完成されたデータを保存または公開し，その後に想定されるさまざまな研究に生かして行くことを念頭に作成されるという形もとられている。

こうして収集された学習者言語データが電子化され，アクセス可能な情報源として保存されたものが学習者言語コーパスである。

本稿は現在公開されている日本語学習者言語コーパスについて調査し，今後の学習者言語研究および日本語教育研究における有効性を目指し記すものである。

1. 先行研究

1.1　コーパスの定義

コーパスという言葉はラテン語の corpus （「体，かたまり＝body」の意）から来ており，英語の corpus は「本体，（文書などの）集大成，言語資料」を意味する語である。言語学においては特定の言語資料（コーパス）を設定し，それについて分析するという研究手法でこの語が使用されてきた。その後コンピュータの発達とともに，データ処理にコン

ピュータが導入され，多量のデータについて語彙索引の作成・検索や統計処理が可能となった。

後藤（1995）ではコーパスを「主にコンピュータによる処理を前提とした機械可読のテキスト，電子（化）テキストの大規模な集合」としている。

研究の対象として多くのサンプルを必要とする場合でも，電子化されたデータは網羅的に処理することができ，言語サンプルとしての位置づけは高い。そればかりでなく，目的に沿ってタグをつける（情報を付加する）ことができる。たとえば学習者言語に見られる誤用を，一定の認定基準に従って分析・分類し，記号化したものをタグとして付加する場合などがこれに当たる。

1.2　言語学研究におけるコーパス利用の歴史

言語資料としてのコーパスは主に英語圏で開発され利用されてきた。利用目的は当初，語法・文法研究，辞書編纂などであったが，その後，非母語話者による誤用や学習者言語に視点が置かれ，英語教育の分野で第二言語習得研究に使用されるようにもなった。

英語圏のコーパスの主なものとしては「ブラウン・コーパス」，Cobuild の辞書・語法書の編纂に利用された「Bank of English」〈語数　4 億 5 千万〉，「英文法書」〈語数　4 万〉（Biber et al., 1999），「British National Corpus（BNC）」〈語数　1 億〉などが挙げられる。「BNC」は特にバランスのとれたコーパスという見方ができ，データの網羅性や構成内容を考える上でひとつの指針を提示している。

外国語教育の視座から学習者言語に焦点を当てた研究は日本国内でも多くの労を傾けて行われている。一例として日本語を母語とする英語学習者については学齢期の生徒を対象としたものから成人学習者まで，学習者言語研究の先導的位置にあって積極的な取り組みがなされている。

1.3　日本語のコーパス研究

日本語のコーパス研究についても，活発な展開を見ることができる。それらを概観した文献により情報入手の方法を知ることも可能である。

松本・小磯（1996）では，研究利用が可能な日本語のコーパスを紹介し，日本国内における言語データ収集のため長期的に活動する組織の必要性を指摘している。その後も国立国語研究所あるいは各種プロジェクトを中心に日本語コーパスが構築されてきた。たとえば新聞記事を文字化し電子化したものや，電子辞書編纂の形で構築された日本語コーパスもある。また，文字化された資料を検索するためのツールの開発も進められている。日本の小説を文字化した「青空文庫」のようなコーパスが日本語研究に果たす役割も大きい。

そうした日本語コーパスは母語としての日本語を文字化した資料であるが，一方，日本語の学習者言語コーパスは学習過程であらわれた「中間言語」（Selinker 1972）と位置づけられるものである。

日本語の学習者言語コーパスについて，上村（1997）は具体的な研究成果として上村コーパスの内容を挙げ，紹介している。

コーパスによる言語研究が進むにつれ，その長所を含めさまざまな特徴が明らかになっている。松田（2001）はコーパスを使用した研究の特性について「コーパスで見つかるものと見つからないもの」という視点から詳述し，その限界を提示している。そして「大量のデータを与えてくれるコーパスが入手できたとしても，研究における問題発見や仮説設定はあくまで研究者の仕事である」とし，コンピュータ利用によって得た時間を価値ある企図や判断に向けるよう鼓舞している。

大曽・滝沢（2003）は言語研究をする上でのコーパスの有効性について，具体例を挙げながら論じている。特に「母語の記述を行うには，コーパス利用の他に，内省に依拠するという方策がある」，すなわち数多く示された例により数量的多数決を取ることの他に母語話者の直観によって正誤や慣例を判断することもできることを認めた上で，次の視点を提示する。「学習者が産出する言語を記述しようとする場合には，研究者が内省を働かせることは不可能であるから，研究者側は学習者の言語を観察して記録することが必要となる」（大曽・滝沢 2003）としている。

このように日本語コーパスの中でも学習者言語コーパスは他言語における位置づけと同様，明確な必要性を持つものとして捉えられ得る。母語ではない日本語を観察し分析することは，日本語教育の上で必要であると同時に日本語そのものの研究にも資するところが大きいと言える。

現在インターネットでアクセス可能であり研究目的にも利用できる日本語の学習者言語コーパスについて，近年の注目すべき研究成果としてその足跡をたどることのできる 4 つのものを次の章で紹介してゆく。

2. 日本語の学習言語コーパス

2.1 KY コーパス

KY コーパスとは，平成 8 年度から平成 10 年度にかけて行われた文部省科学研究費助成プロジェクト「第 2 言語としての日本語の習得に関する総合研究」の成果の一部である。内容は OPI のインタビューをデータとしており，外国人学習者の発話資料から成っている。

OPI（Oral Proficiency Interview）は ACTFL（American Council on the Teaching of Foreign Languages，全米外国語協会）が実施しているもので口頭能力を測定するための面接テストである。15 分から 30 分の面接によって，話す技能をさまざまな角度から観察する。レベルは初級（上中下）・中級（上中下）・上級（上中下）・超級の 10 段階に分けられ，テスト結果を分析し評価する。

KY コーパスは 90 人分の OPI テープを文字化した言語資料である。被験者を母語別に見ると，中国語・英語・韓国語がそれぞれ 30 人ずつとなっている。さらに 30 人の OPI 判定結果別の内訳は，それぞれ初級 5 人，中級 10 人，上級 10 人，超級 5 人となっている。

KY コーパスの K と Y はコーパス作成を担当した鎌田修（南山大学人文学部教授）・山内博之（実践女子大学助教授）両氏の頭文字である。

KY コーパスの特長すなわち OPI のデータを言語コーパスとして用いることの利点につ

いて，作成者自ら次のように記している。

第一に各被験者の能力レベルが明示されていることが挙げられる。学習歴や在日歴，または在籍クラスのレベルといったあいまいさを含む基準ではなく，OPI のテストにおける「初級―上」「中級―中」等の能力が示されていることは，第二言語習得研究に使用する際非常に有効である。

第二にデータ採取の手法が標準化されていることである。インタビューの形式が同レベル内で統一されているため，データ同士の比較がしやすく，分析が容易である。

第三に発話単位の認定が比較的容易であり，数量化，定量化を行いやすい。これは OPI のインタビューが基本的に「質問・応答」の繰り返しという形をとっているため，被験者の応答の部分を一発話単位として認定できるということである。

2.2　上村コーパス

上村コーパスはインタビュー形式による日本語会話データベースである。

OPI テスターが日本語母語話者 54 名，非母語話者 66 名，計 120 名に行った日本語 OPI の文字化テキストを収録したもので，上村隆一氏（北九州市立大学，当時）他の編集によるものである。

日本語母語話者と非母語話者の発話パターンの比較分析と，日本語教育向けの基礎資料としての日本語会話コーパス構築を目的としている。

1991 年度から準備研究を始め，1995 年以降は科研プロジェクト「人文科学とコンピュータ」の一環としてコーパス作成を行った。1998 年に CD-ROM『インタビュー形式による日本語会話データベース』として完成している。

データの収録は日本国内およびアメリカで行われた。被験者は日本語母語話者である大学生，大学教職員，日本語教師，主婦，会社員などと，非母語話者の留学生，就学生，大学教員，会社員などである。

種類	目的・内容	データ収集地	データ収集の対象			人数または データ数	データ資源	特徴	作成時期・ プロジェクト名	作成者・ 研究代表者
			日本語の レベル	出身地、 または母語	社会的身分					
KYコーパス	日本語学習者の発話資料を元にした言語コーパスの構築（次の特長を持つ） 1）能力レベルを明示 2）データ同士の比較が容易 3）発話単位の数量化・定量化が容易		初級（5） 中級（10） 上級（10） 超級（5）	中国語（30） 英語（30） 韓国語（30）			90 テープ録音を文字化	ACTFLのOPIに準拠 ・各被験者のproficiency による能力レベルの明示 ・データ採取の手法の標準化 ・発話単位の認定が容易 （「質問→応答」の形式）	1996年度～1998年度 科学研究費補助研究 「第2言語としての 日本語の習得に関する 総合研究」成果の一部	鎌田修 （南山大学） 山内博之 （実践女子大学）
上村コーパス （日本語会話 コーパス プロジェクト）	日本語母語話者と非日本語母語話者の発話パターンの比較分析と日本語教育向けの基礎資料となる日本語会話コーパス構築	アメリカ （大学　1） 日本 （大学　1） （日本語学校　1）		アメリカ(27)韓国（22） 中国(2)台湾（2） デンマーク(2)ロシア(2) オーストリア(1) オーストラリア(1) シンガポール(1)タイ（1） デンマーク(1)ドイツ(1) トリニダード・トバゴ(1) バングラデシュ(1) フィリピン(1)	（非日本語 母語話者） 学生 大学教員 会社員、等 （日本語 母語話者） 大学教員 日本語教師 会社員、主婦、等	（非日本語 母語話者） （日本語 母語話者） 54	インタビュー実験による会話・ 66 ロールプレイ （音声・文字化資料）	ACTFLのOPIに準拠 ・データの形式と内容の一貫性	1991年度～準備研究 1995年度～1996年度 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	上村隆一 （北九州市立大学）
名古屋大学 作文コーパス	日本語教育、日本語学、言語学、情報処理に有益な学習者の作文コーパスの構築 [具体的内容] 1）一定量の作文を収集、電子化 2）誤りの訂正を加える 3）誤りの原因、種類を分析して加える 4）完成した作文コーパスを公開	日本 （大学　2） （電子メール）	初級～上級	英語、中国語、他	学生	756ファイル	作文、電子メール、課題レポート他	（入力） CHILDSのCHATフォーマットを使用 （分析） 誤用の原因、種類のタグ付け （フォーマットチェック） checkerプログラムを作成、確認 （検索ツール） 検索ツールを作成	1996年度～1998年度 科学研究費補助基盤研究 「日本語学習者の 作文コーパス: 電子化による共有資源化」	大曽美恵子 （名古屋大学）
日本語学習者による 日本語作文と、 その母語訳との 対訳データベース	1）日本語学習者が書いた日本語作文 （400～800字） 2）執筆者本人による1）の母語訳 （またはもっとも楽に文章を書ける言語への翻訳） 3）日本語教師による作文の添削 4）作文執筆者・添削者の言語的履歴に関する情報 以上4種類の収集、データベース化	中国、インド、 カンボジア、韓国、 マレーシア、 モンゴル、 シンガポール、 ベトナム、日本		中国、インド、 カンボジア、韓国、 マレーシア、モンゴル、 シンガポール、 ベトナム、日本			1100 作文、母語訳	（作文）テクストファイル pdf 画像(jpg)ファイル （母語訳）テクストファイル または MS-Wordファイル pdf	1999年度～2000年度 科学研究費補助基盤研究 「日本語教育のための アジア諸言語の対訳作文 データの収集とコーパス の構築」 2000年度 科学研究費補助研究 「日本語学習者による 日本語作文とその母語訳 との対訳コーパス」	前田（宇佐美）洋 （国立国語研究所）

研究チームメンバーは代表者：上村隆一（福岡工業大学），田吹昌俊（九州工業大学）根津真知子（国際基督教大学），村野良子（国際基督教大学），横田将生（福岡工業大学）の各氏である。

会話コーパスの形式はインタビュアーによる個人インタビューである。時間はおよそ 15 分から 30 分で，会話モードとロールプレイモードから成っている。インタビューの形式は OPI に沿ったものであるが，OPI のレベルチェックは行われていない。

2.3 名古屋大学作文コーパス

日本語研究，日本語教育の推進に有用な電子化資料（コーパス）の収集・作成を目指し，名古屋大学国際言語文化研究科から発信したプロジェクトとして公開を念頭に置き大規模な学習者言語コーパスの構築を行っている。

ここに挙げるのは関西外国語大学および名古屋大学で学ぶ留学生の作文やアメリカの学生の日本人宛メールといった書き言葉の資料を文字化し，電子化したものである。入力には CHILDES の CHAT フォーマットを使用している。

収集した資料にはさらに誤用についての情報がタグとして付けられている。

まず個々の誤用について益岡・田窪（1992）による文法規準の分類に従い分析を行う。そして誤用の範囲や分析結果を記している。

20 カ国に及ぶ国籍の日本語学習者の書き言葉の資料に，分析済みの情報を付加したタグ付データという有益な情報源を公開し，広く提供する形となった。

2.4 日本語学習者による日本語作文と，その母語訳との対訳データベース（国立国語研究所）

このデータベースは平成 11〜12 年度科学研究費補助金基盤研究（B）(2)「日本語教育のためのアジア諸言語の対訳作文データの収集とコーパスの構築」（課題番号：国 11691041，研究代表者：前田（宇佐美）洋）ならびに，平成 12 年度科学研究費補助金（研究成果公開促進費・データベース）「日本語学習者による日本語作文とその母語訳との対訳コーパス」（課題番号：128049，研究代表者：前田（宇佐美）洋）の研究成果として作成されたものである。

1999 年度から 2000 年度にかけてアジア 10 か国（中国・インド・カンボジア・韓国・マレーシア・モンゴル・シンガポール・タイ・ヴェトナム・日本）から 1100 名分のデータを集めた。データベースは次の内容から成っている。

1. 日本語学習者による日本語作文
2. 作文執筆者本人による母語訳（またはもっとも楽に文章を書ける言語への翻訳）
3. 日本語教師による作文の添削
4. 作文執筆者・添削者の言語的履歴に関する情報

以上 4 種類のものを大量に集め電子化した上で検索のためのインデックスを付けた。比較のため日本語母語話者の作文も収集した。

2001 年 3 月，CD-ROM の形で公開され，さらにその後 2004 年 3 月，2005 年 4 月にデー

タの追加が行われ，コンピュータネットワーク上で公開されている。このデータベースはユーザ登録した上での利用が可能である。

3. コーパスを活用した論文

学習者言語コーパスは主として研究目的の利用を目指し構築されてきた。

名古屋大学作文コーパスでは収集されたデータを起点にさまざまなプロジェクトが企画され，各研究成果として報告論文集の形で発表されている。まず『日本語学習者の作文コーパス：電子化による共有資源化』[平成 8 年度～10 年度科学研究費補助金（基盤研究 A (1)）研究成果報告書（研究課題番号 08558020）研究代表者：大曽美恵子（名古屋大学大学院国際言語文化研究科教授）]がある。

ここでは学習者の作文を文字化する際の入力マニュアルや誤用に関するタグを付した際の誤用認定の基準・方法を示すと同時に，コーパスを利用した研究すなわち「て形の習得順序」や「コロケーションの誤用」をはじめとする研究成果としての論文が掲載されている。

さらに『日本語電子化資料収集・作成―コーパスに基づく日本語研究と日本語教育への応用を目指して―』[平成 12 年度名古屋大学教育研究改革・改善プロジェクト報告論文集]が挙げられる。この中でたとえば杉浦（2001）は「コーパスを利用した日本語学習者と母語話者のコロケーション知識に関する研究」と題する研究成果を発表している。またその後の『日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究』[平成 13 年度科学研究費補助金基盤研究（B）(2)（研究課題番号 13480069）中間報告論文集]では杉浦・朴 (2003)「日本語学習者作文コーパスにおける形態素レベルでの共起表現について」と題する研究が展開されている。上記の 2 つの論文とも名古屋大学作文コーパスの資料を活用したものであり，コーパス資料に誤用分析を目的とする誤用タグが付けられているため，このタグを取り外し学習者言語コーパスとして分析に用いている。学習者言語データと母語話者によるデータが比較検討されており，こうした形での活用の有効性も特徴的である。

KY コーパスや上村コーパスについても，研究に資するところは大きく，たとえば「KY コーパスを利用した論文」という形で労作の数々が発表されてきた。松田（2001）は KY コーパスにおける韓国語母語話者の日本語発話について分析し，「を」格に関する文法的特性に注目し論じている。さらに松田（2002）「中間言語と言語変異：KY コーパスを使った「を」格ゼロマーク化の分析」において研究の発展を示している。中石（2005）は KY コーパスから対のある自他動詞を取り出し，使用状況の分析から規則性を探っている。このように第二言語習得研究会等の場を中心に，コーパス活用の有意性が証明されつつ現在に至っている。

「日本語学習者による日本語作文と，その母語訳との対訳データベース」については，たとえば『日本語教育のためのアジア諸言語の対訳作文の収集とコーパスの構築』(2000) にその成果報告を見ることができる。

4. 今後の展望

4.1 コーパスの使用

日本語コーパスを研究用に使用するための自動解析ツールは各所で開発されている。コーパスの中には付加情報を加えたものや自動解析にかけた後に人手による修正を加えたものなどタグ付きコーパスとして公開されているものもある。また一方でコーパスを入手し自動解析ツールを使用する方法も可能である。さらに，コーパス自体は公開されていなくても登録後に語単位で検索し情報が得られるようなサービスも確立しつつある。

コーパスの使用に当たっては有償・無償を問わず作成・管理者との間の使用許諾契約や引用の際の著作権の問題など，配慮すべき点が含まれる。

学習者言語という特徴を生かしたコーパス使用では，日本語学習支援ツールの開発が挙げられる。赤堀（1996）で紹介されている学習支援ツールは，学習者が受身についての文を入力すると誤りの箇所が訂正されるというものである。このシステムを可能にしているのは，学習者の作文から受身に関する誤りを調査・分類し，その分類に基づいて作成された誤り処理のルールである。すなわち学習者が産出する可能性のある誤りの項目があらかじめ設定されている。こうして，日本語を学習しようとする者が自由に入力した文に対して，誤りの検出と適切なフィードバックがなされる。

杉浦（1999）が提示している学習支援ツールは，誤用分析がなされた学習者言語コーパスを利用する語法検索ツールと作文推敲ツールである。学習者の作文の誤用を収集し分析したコーパスを背景として持つことにより，学習途上にある者の作文について誤りの箇所を指摘し，間違えやすい項目や誤用例を示すというシステムである。これらの領域でも，コーパス活用が支える研究発展の可能性は大きい。

4.2 展望

今後，日本語学や日本語教育，第二言語習得研究といった分野において学習者言語コーパスや日本語のコーパスの果たす役割はますます重要視されることと思われる。しかしながらコーパス作成に際しては資料提供者について匿名・仮名その他個人が特定されないなどの配慮が必要となる。従ってそうした状況に対処するための手続きを行う作成者側の腐心は収集自体の労力とあいまって計り知れないものがある。またデータの量はもちろん条件の整備・情報内容のバランス等，質的な面の充実についても考慮されねばならない。多くの人々の協力を得，幾段階もの作業を経ることになるが，それらに鑑みてなお研究の発展に資するコーパスの開発が望まれる。また開発されたコーパスの特徴を最大限に生かした活発な理論の展開がなされてゆくよう，今後の更なる発展を期待したい。

参考文献

赤堀侃司（1996）「自然言語処理を用いた日本語作文学習支援システムの開発」『日本教育工学会研究報告集』pp.61-68　日本教育工学会

上村隆一（1997）「データベースで調べる」『日本語学』11 pp.60-68　明治書院

大曽美恵子（1999）『日本語学習者の作文コーパス：電子化による共有資源化』平成 8 年度～10 年度科学研究費補助金基盤研究（A）(1) 課題番号 08558020　研究成果報告書　名古屋大学

大曽美恵子・滝沢直宏（2003）「コーパスによる日本語教育の研究―コロケーション及びその誤用を中心に―」『日本語学』22（5）　pp.234-244

国立国語研究所（2000）『日本語教育のためのアジア諸言語の対訳作文とコーパスの構築』平成 11 年度～12 年度科学研究費補助金基盤研究（B）(2) 研究成果報告書

後藤斉（1995）「言語研究のデータとしてのコーパスの概念について―日本語のコーパス言語学のために―」『東北大学言語学論集』第 4 号 pp.71-87

迫田久美子（1998）『中間言語研究』渓水社

杉浦正利（1999）「誤用データの検索と日本語学習者支援ツールの開発」『日本語学習者の作文コーパス：電子化による共有資源化』平成 8 年度～10 年度科学研究費補助金基盤研究（A）(1) 課題番号 08558020　研究成果報告書　名古屋大学

杉浦正利（2001）「コーパスを利用した日本語学習者と母語話者のコロケーション知識に関する調査」『日本語電子化資料収集・作成―コーパスに基づく日本語研究と日本語教育への応用を目指して―』平成 12 年度名古屋大学教育研究改革・改善プロジェクト報告論文集 pp.64-81　名古屋大学国際言語文化研究科

杉浦正利（2002）「コーパスに基づいた外国語作文支援システム」『日本語学と言語教育』東京大学出版会

杉浦正利・朴秀智（2003）「日本語学習者作文コーパスにおける形態素レベルでの共起表現について」『日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究』平成 13 年度～平成 15 年度科学研究費助成金基盤研究（B）(2)（研究課題番号 13480069）中間報告論文集 pp.1-10　名古屋大学国際言語文化研究科

中郷慶（1999）「コーパス言語学の現状と課題」『愛知淑徳短期大学研究紀要 38』

中石ゆうこ（2005）「対のある自動詞・他動詞の第二言語習得研究―「つく-つける」,「きまる-きめる」,「かわる-かえる」の使用状況をもとに―」『日本語教育』124 号　日本語教育学会

成田真澄（2004）「コーパスに基づく第二言語習得研究」『第二言語習得の現在―これからの外国語教育への視点』大修館書店

益岡隆志・田窪行則（1992）『基礎日本語文法』くろしお出版

松田謙次郎（2001）「中間言語と言語変異：KY コーパスを使った「を」格ゼロマーク化の分析」『Theoretical and applied linguistics at Kobe Shoin: トークス 4』pp.57-76　神戸松蔭女子大学

松田謙次郎（2001）「コーパス調査と計量的研究」『日本語学』20（5）　pp.32-41

明治書院
松本裕治（2003）「現代語のコーパスの種類とそれぞれの特徴」『日本語学』4 月臨時増刊号 pp.54-60 明治書院
松本裕治・小磯花絵（1996）「電子化時代の言語コーパス 1　日本語のコーパス」『言語』25（10）　pp.114-120　大修館
Selinker,L.（1972） Interlanguage. *International Review of Applied Linguistics*, 10, 209-231