Learners' Spoken Corpus of Japanese and Developmental Sequence of Verbs

国語研プロジェクトレビュー
NINJAL Project Review　Vol.3 No.3 pp.107–116（March 2013）

〈共同研究プロジェクト紹介〉
多文化共生社会における日本語教育研究
サブプロジェクト：学習者の言語環境と日本語の習得過程に関する研究

日本語学習者の発話コーパスと動詞の発達

Learners' Spoken Corpus of Japanese and Developmental Sequence of Verbs

迫田 久美子（SAKODA Kumiko）

1. はじめに

第二言語習得研究とは，目標とする外国語を学習・習得する過程で起きるさまざまな現象をとりあげ，科学的に研究し解明することである。第二言語習得研究の歴史は浅く，国内や海外の日本語教育の広がりと共に，論文数は徐々に増えてきたが，第二言語として日本語を習得する学習者のデータは極めて少ない。その背景には，個人での研究ではデータ収集が困難であり，時間がかかるなどの理由がある。しかし，言語習得の研究には，学習者の言語データは不可欠である。

本稿では，「学習者の言語環境と日本語の習得過程に関する研究」のサブプロジェクトが日本語学習者の縦断的発話コーパス（C-JAS）を開発した過程を示し，そのコーパスの動詞分析から学習者の日本語習得過程の一端を明らかにすることを目的とする。本稿では，コーパスを「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。

2. 日本語学習者の発話コーパス

2.1 これまでの発話コーパス

これまでに公開された日本語学習者の発話コーパスを概観し，現段階における問題点を挙げる。次ページの表 1 は，これまでに公開された日本語学習者の発話コーパスの一覧表である。表 1 の横断調査とは，多数の調査対象者に対して一斉に調査を実施する場合であり，縦断調査とは，特定の調査対象者に対して，ある程度の長い期間にわたり，その変化を調査する場合である。

8 件のデータベースのうち，(1)～(3) の 3 件は OPI（Oral Proficiency Interview）という口頭能力インタビュー試験を用いており，学習者のレベルが初級～超級まで 9～10 段階で判定されている。

(4) は，社会言語学的観点から，BTSJ（Basic Transcription System for Japanese）という記述方法によって，さまざまなグループ（日本語母語話者同士，日本語母語話者対学習者，親しい者同士・初対面者同士等）における，さまざまな場面（雑談，論文指導，電話等）の対話を収集している。

(8) は，台湾の東呉大学の 1 年生 37 名を 3 年半にわたって調査した縦断データである。

60 分間で作文を書かせ（作文 1），その後，日本語教師が学習者に 1 対 1 でフォローアップインタビューを 20 分間行い，その後，再び修正文（作文 2）を作成する。作文は，作文コーパスとして公開し，口頭のフォローアップインタビューを発話コーパスとしている。

表 1　日本語学習者を対象とした日本語の発話コーパス一覧（網掛けは，縦断調査を示す）

	名　称（作成者）	a）学習者の母語　b）レベルと学習者数　c）収集方法　d）URL
(1)	KY コーパス （鎌田修・山内博之）	a）英語・韓国語・中国語 b）初級～超級レベルの学習者　90 名 c）口頭能力インタビュー試験（OPI）による 30 分の対話（横断調査） d）http://opi.jp/shiryo/ky_corp.html（説明） （タグ付き　http://www30.atwiki.jp/corpus-ling/pages/57.html）
(2)	日本語学習者会話 データベース （国立国語研究所）	a）英語・韓国語・中国語・インドネシア語・その他 b）初級～超級レベルの学習者　339 名 c）口頭能力インタビュー試験（OPI）による 30 分の対話（横断調査） d）https://dbms.ninjal.ac.jp/nknet/ndata/opi/
(3)	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）	a）韓国語・中国語・ロシア語・タガログ語・ポルトガル語・その他 b）初級～超級レベルの学習者　25 名 c）OPI の枠組を活用したインタビューによる 30 分の対話（縦断調査 1～2 年） d）https://dbms.ninjal.ac.jp/judan_db/
(4)	BTS による多言語話し言葉コーパス （宇佐美まゆみ）	a）韓国語・中国語 b）294 会話 c）日本語母語話者同士，日本語母語話者と非母語話者の対話など（横断調査） d）http://www.tufs.ac.jp/ts/personal/usamiken/index.htm（説明）
(5)	日本語学習者による 日本語/母語発話の対照言語データベース （国立国語研究所）	a）韓国語・中国語・タイ語 b）学習者　190 名，日本語母語話者　57 名 c）スピーチとロールプレイによる発話（横断調査） d）http://jpforlife.jp/hatsuwadb.html
(6)	初対面日本語会話 データベース （伊集院郁子）	a）中国語 b）学習者　4 名，日本語母語話者　4 名 c）日本語母語話者同士，日本語母語話者と上級レベル以上の学習者との会話（横断調査） d）http://www.tufs.ac.jp/ts/personal/ijuin/koukai_data1.html
(7)	日本語学習者会話 ストラテジーデータ （国立国語研究所）	a）英語・その他 b）学習者　10 名 c）日本語母語話者（店員や係員）との 1 対 1 の会話（横断調査） d）https://dbms.ninjal.ac.jp/nknet/ndata/strategy/
(8)	LARP at SCU （Language Acquisition Research Project at Soochow University） （台湾東呉大学）	a）中国語 b）学習者　37 名 c）毎月 1 回の作文データ収集時の発話による内省データ（縦断調査 3 年半） d）http://webbuilder.scu.edu.tw/builder/web_page.php?web=156&pid=9346

2.2 発話コーパスの課題

これまでの発話コーパスは，表 1 からもわかるように（8 件中，縦断調査が 2 件で，6 件は横断調査），縦断調査のデータが少ない。また，(3) は，データ収集が 1 年に 1 回であるため，学習者 1 人に関して 2 回分（60 分）のデータしかない。このため，縦断的な変化を分析することは難しい。

さらに，(8) では，母語が中国語のみであり，作文作成直後のフォローアップインタビューであるため，話題が作文に限られてしまい，一般的な話題による会話データとするには，問題が残る。また，8 件中，形態素解析や誤用抽出ができるシステムが備わっているのは (1) のみで，多くのコーパスには備わっていない。

以上から，次の 3 点が現段階での日本語学習者の発話コーパスの問題点として挙げられる。

(9) 縦断調査のコーパスが少ない。
(10) 既存の縦断調査のコーパスでは，量が少なく，話題の範囲も狭い。
(11) システム検索ができるコーパスが少ない。

3. 日本語学習者の縦断的発話コーパスの開発

先述の発話コーパスの問題点をふまえ，3 年間の縦断調査による発話コーパス（C-JAS Corpus of Japanese as a Second Language: 第二言語としての日本語のコーパス）の開発について述べる。

3.1 縦断的発話コーパス（C-JAS）の特徴

「学習者の言語環境と日本語の習得過程に関する研究」のサブプロジェクトは，縦断的発話コーパス（C-JAS）（以下，本コーパス）に関して，「第二言語習得研究の普及」「言語研究への貢献」「日本語教育への応用」の 3 点を目的として開発を開始した。2. で指摘したように，縦断調査と言っても 1 年に 1 回のデータ収集ではその量は極めて少ない。第二言語習得研究を展開するには，母語を統制した特定の学習者の長期にわたるデータが必要である。そのような日本語のデータは，分析によって言語の発達過程や習得要因などが明らかになり，言語学の研究領域でも貴重な資料となる。さらに，それらの知見は日本語の指導へと応用することが期待できる。

サブプロジェクトで開発する本コーパスの特徴は，以下の 3 点である。

(12) 文法項目や談話表現の習得に関する研究を主な目的としている。
(13) 中国語話者と韓国語話者の 3 年間の縦断的発話データである。
(14) 全文閲覧機能と検索システムを備えている。

3.2 本コーパスの概要

本コーパスは，18 歳～25 歳の同一日本語学校に在籍した中国語母語話者 3 名（C1～3,

女性）と韓国語母語話者 3 名（K1〜3，女性 1 名，男性 2 名）による学習開始 3 ヵ月から約 3 年間の発話をデータとしている。学習者は全員同じ日本語学校で初級から学んだ教室環境学習者であり，その際に使用された教科書は『日本語初歩』（国際交流基金編 1985）であった。

データ収集の時期は，1991 年 7 月〜1994 年 3 月で，1991 年 4 月に学習を開始してから 3〜4 ヵ月に 1 回の割合でデータ収集を行い，3 年間で学習者ごとに 7〜8 回の発話データを得た。1 回の調査のデータは，約 60 分で，本コーパスの長さは約 46 時間 30 分である。

3.3　形態素タグの付与

本コーパスは検索の利便性を向上させるため，一般的な文字列検索だけでなく，形態素情報を用いた検索が行えるように，文字化データに対し，形態素解析を行った。形態素解析とは，コンピュータを用いて文を形態素に分割し，それぞれの品詞を判別する作業のことを言う。本コーパスでは，MeCab と UniDic を使用している。

形態素解析の例を挙げて説明する。「不安でしょ」という発話には，表 2 のような形態素情報が付与される。表 2 の項目で，「語彙素」とは，国語辞典の見出し語に相当する語のことを指す。

表 2　形態素解析の例

書字形	発音形	語彙素読み	語彙素	品　詞	活用型	活用形	語形	語種
不安	ふあん	ふあん	不安	名詞・普通名詞 形状詞可能			フアン	漢
でしょ	でしょ	です	です	助動詞	助動詞 デス	意志推量形		和

検索に形態素情報を活用することにより，「多様な活用形を一度に検索できる」「品詞情報を利用して対象とする語を検索できる」などの利点が生じる。たとえば，前者では，「書く」という語彙素をキーワードとして検索することで，「書かない」「書きたい」「書いた」のようなすべての活用形を一度に検索することが可能となる。また，後者では，助詞の「は」を単独で検索する場合，文字列検索では「こんにちは」のような不適切な例が大量に抽出されるが，形態素解析されたデータであれば，品詞を「助詞」と指定して検索することで，不適切な例を除外することが可能となる。

しかし，音声言語から成るデータを形態素解析する場合，複数の技術的な問題が生じる（李 2009）。省略や言い直しが多量に含まれるので，多くの誤解析や，本来の意図と異なる解析結果が発生してしまう。そのため，それらを修正し，解析精度を高めていくのは本コーパス構築でも重要な作業である。

3.4　誤用タグの付与

形態素情報の付与と異なり，誤用情報の付与には誤用判定の主観が大きく関わる。また，研究目的によって付与の方針も変わってくる。作文では，多くのコーパスで誤用タグが付与され，分類されているのに対し，発話コーパスでは，KY コーパスを除くと，ほとんど付与されていない。

発話コーパスに誤用タグが付与されない理由としては，発話では，省略や倒置などが起こり，文や発話の区切りを明確にとらえることができず，誤用の判断が極めて難しい点が挙げられる。また，誤用の判定も個人や研究目的によって変わってくるため，統一的に誤用のタグをつけることは，多大な困難が伴う。

本コーパスでは，誤用の判定については慎重に行いたいので，まずは，誤用の有無についてタグをつけ，誤用の分類については，検討を重ねた上で付与する方針を立てている。つまり，どの種の誤用であるかは不明であるが，誤用表示のタグをつけ，わかる範囲内で正用を付加することとした。したがって，誤用の個所とその正用についての情報を提供することができる。

これらのタグを付与する上で想定した利点は，「検索対象を増やすこと」，「誤用個所を判別しやすくすること」の 2 点である。前者は，主に「発音」の誤用に関連しており，発音が誤っている場合は，文字化データの形態素解析の過程で正確な形態素情報が付与されない可能性が高い。後者に関しては，検索過程で誤用個所が判別しやすくなることを想定した。

4.　本コーパスにみる動詞の発達過程

4.1　L1 と L2 の類似点

本コーパスは，学習開始から 3〜4 ヵ月に 1 回の割合でデータが収集された。したがって，3 年間の習得過程は，8 つの時期に分けて観察できる。日本語学習者は，活用形をどのように発達させるのだろうか。図 1 は，使用頻度の高かった動詞「思う」を取りあげ，各時期に韓国語話者 K1 の発話にどのような語形で新しく現れたかを調べ，新出語形を示したものである。

形式が発達していく過程で，核となる形式を太字で示し，発展した形式を矢印で示した。たとえば，新出語形（第 3 期「**思います**よね」「**思います**けど/から」第 4 期「**思います**が」）がそれ以前に出現した語形（第 2 期「**思います**」）に，形態素（「よね」「けど/から」「が」）を加えた言語形式である場合は，古い語形「**思います**」から長くなった新出語形「**思います**よね」「**思います**けど/から」「**思います**が」へ矢印「↓」を記入した。

K1 は，日本での滞在が長くなると，既出の動詞語形にさまざまな要素を付け加え，新出語形を形成している。たとえば，第 3 期に「**思う**」を使い，第 4 期には「から」が付加されて，「**思う**から」が出現し，第 5 期には「**思う**ですよね」が見られた。また，第 3 期の「**思った**」は，第 5 期には「**思った**です/ですが/ですよ/ですけど」と広がり，第 6 期には「**思った**ですね」が観察された。

このような活用が長くのびていく現象は，日本語母語話者の幼児の言語にも観察される。

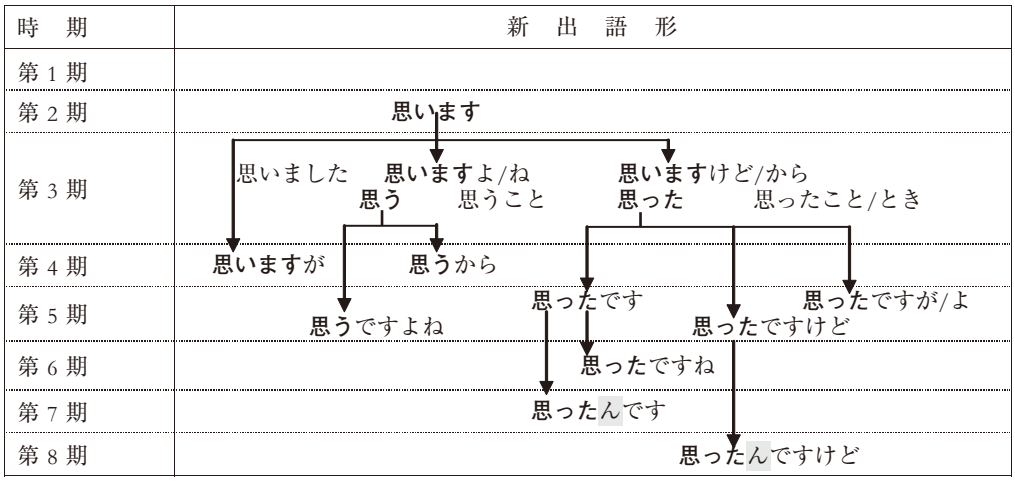

図 1　時期別の韓国語話者 K1 の動詞「思う」の新出語形（一部抜粋）（迫田 2012: 118 修正）

岩立（1981）は，2 歳児のデータで，古い形に新しい要素が付加されて動詞が発達していく過程を明らかにしている。たとえば，動詞「食べる」は，（15）のように発達が進む（「2;1」は「2 歳 1 ヵ月」を表す）。

（15）　**タベル**（2;1）　→　**タベル**ノ（2;3）　→　**タベル**ノヨ（2;5）
　　タベタイ（2;3）　→　**タベタイ**ノ(2;4)/**タベタイ**ナ(2;5)/**タベタイ**ヨ(2;7)
　　タベタ（2;3）　→　**タベタ**ノ（2;8）

岩立（1981）は，このように古い動詞形に新しい形態素の要素が付加されて動詞が発達するとする考えを「くっつき仮説」として発表した。本コーパスの日本語学習者の習得過程においても，図 1 に見られるように同様の現象が観察された。このことは，第一言語（L1）と第二言語（L2）の習得過程が類似していることを主張する証拠となる。

4.2　L1 と L2 の相違点

では，L1 と L2 の習得過程は同じなのであろうか。表 3 は，日本人幼児と日本語学習者の動詞「食べる」の初出の活用形を示したものである。どちらも，初出形に多様な形式が見られ，日本人幼児は，2 歳 1 ヵ月に「タベル・タベチャッタ・タベナイ」が観察されている。一方, 学習者の場合は,「食べる」の使用は, 第 1 期（学習 3〜4 ヵ月）から第 3 期（学習 1 年）の間に見られ，その初出形には日本人幼児とは異なった形式が観察される。

学習者には「**タベマス**」「**タベマスヨ**」「**タベルンデス**」などの丁寧体が出現している点が特徴的である。本コーパスの学習者は, 教科書（『日本語初歩』国際交流基金編 1985）によって動詞は最初に「です・ます」の丁寧体が教えられている。初出形に丁寧体が多く見られることに教科書からの何らかの影響があることが推測される。

表 3　日本人幼児と日本語学習者の「食べる」の出現時期と初出形

	幼児 (2;1)	K1 第 3 期	K2 第 2 期	K3 第 2 期	C1 第 1 期	C2 第 1 期	C3 第 3 期
初出形	タベル タベチャッタ タベナイ	タベタラ	タベタ，タベタラ タベタイノ/モノ タベテ/テモ/テネ **タベタインデス**	タベルカラ タベテ **タベルンデス**	タベナイ	**タベマス**	タベタリ **タベマス** **タベマスヨ** タベナサイ

4.3　中間言語形の発達

図 1 の「思う」の新出語形には，規範的でない語形が見られる。第 5 期に出現する「思うです」や「思ったです」で，具体的には（16）のような発話である。このような「動詞普通体＋です」の誤用は，「思う」の動詞だけでなく，K1 の他の動詞にも観察され，さらに，韓国語話者のみならず，（17）のように中国語話者にも観察される。

（16）　日本人と，一緒に勉強することは，本当，むじゅかしいと思ったですよ
（17）　日本は風呂があるですよ（C3 第 7 期）/気を使うですよ（C2 第 5 期）

図 1 では，第 5 期の「**思った**です」は，第 7 期で「**思った**んです」となり，第 5 期の「**思った**ですけど」は第 8 期で「**思った**んですけど」と「のだ」文となっている。このことから，「動詞普通体＋です」の形式は，「動詞普通体＋んです」の中間言語形の一種と考えられないだろうか。

中間言語とは，学習者の母語とも目標言語とも異なった学習者特有の言語体系であり，目標言語への過渡期に出現する形式である。そう考えれば，発達段階の産物であると言える。

他の動詞でもこのような現象が見られるかどうかを調べた。「動詞普通体＋です」の誤用形が出現した時期と「動詞普通体＋んです」の正用形が出現した時期を比較したのが表 4 である。

表 4　誤用形（動詞普通体＋です）と正用形（動詞普通体＋んです）の出現時期の比較

	K1	K2	K3	C1	C2	C3
誤用形(例 違うです)	第 1 期	第 1 期	第 2 期	—	第 2 期	第 3 期
正用形(例 違うんです)	第 7 期	第 2 期	第 1 期	第 7 期	第 2 期	第 5 期

表 4 をみると，1 名は誤用形が出現していないので比較できないが，4 名に関して言えば，正用形の「のだ」文よりも中間言語形の出現が同時か先行している。このことから，誤用形「動詞普通体＋です」は，「動詞普通体＋んです」の前段階の中間言語形の一種であり，「んです」の形式が習得される以前か同じ段階で「動詞普通体＋です」の誤用が産出される可能性が示唆される。しかし，K3 は正用形の出現の後で誤用形が出現している。K3 に正用形が

早く出現したのは，他の学習者に比べるとK3にはホストファミリーがあり，日本人家族との接触も多かったことが影響を与えた可能性が考えられる。

5. おわりに

本稿は，日本語学習者の縦断的発話コーパスを開発した過程を示し，そのコーパスを分析することによって学習者の動詞の発達のプロセスの一端を明らかにすることを目的とした。

本稿の前半では，本コーパスの開発の経緯と特徴を述べた。本コーパスは，従来の発話コーパスの課題であった縦断研究の調査データに基づいており，中国語話者3名，韓国語話者3名の3年間にわたる発話コーパスである。データには，形態素タグ，誤用タグを付与し，検索システムを備えた。

後半では，動詞「思う」の活用形がどのように発話されているのかを調査し，初出形を調べた。そして，古い形に新しい要素が加わって新しい形ができることが明らかになり，それは幼児の第一言語習得と同様の傾向であることがわかった。同時に，成人の日本語習得では，幼児には見られない「です・ます」が早い時期に初出形として観察された。また，「違うです」「思うです」のような誤用は，「違うんです」「思うんです」の前段階の中間言語形として位置づけられることが示唆されたが，断定はできないので，さらなるデータを見ることで，より広く，深く分析を行っていく必要がある。

C-JASは，以下のウェブサイトより利用できる（ユーザー登録が必要）。

https://ninjal-sakoda.sakura.ne.jp/c-jas/web/

●付記●

・本稿で示した日本語学習者の縦断的発話コーパス（C-JAS）の開発は，木下藍子（国立国語研究所），小西円（国立国語研究所），李在鎬（筑波大学）の共同研究である。

・縦断的発話コーパス（C-JAS）は，調査対象となった日本語学習者6名の3年間の協力とそのデータを文字資料化した広島大学大学院の学生達の協力によって基礎資料が作成された。彼らの支援がなくては，本コーパスは完成できなかった。ここに深く感謝の意を表したい。

●参照文献●

岩立志津夫(1981)「一日本語児の動詞形の発達について」『学習院大学文学部研究年報』27: 191−205.

国際交流基金（編）(1985)『日本語初歩』東京：凡人社.

李在鎬(2009)「タグ付き日本語学習者コーパスの開発」『計量国語学』27(2): 60−72.

迫田久美子(2012)「非母語話者の日本語コミュニケーションの工夫」野田尚史（編）『日本語教育のためのコミュニケーション研究』105−124. 東京：くろしお出版.

《要旨》 第二言語習得研究には，学習者の言語データが不可欠である。「学習者の言語環境と日本語の習得過程に関する研究」のサブプロジェクトでは，日本語学習者の言語コーパス，C-JAS を開発した。本稿は，C-JAS の特徴と C-JAS によって観察された動詞の発達について報告するものである。C-JAS の特徴は，中国語母語話者 3 名，韓国語母語話者 3 名の 3 年間の縦断的発話コーパスであり，形態素タグと誤用タグが付与され，システム検索できる点にある。

C-JAS で動詞「思う」と「食べる」の時期ごとの初出形を分析した結果，日本人幼児の第一言語習得と類似した現象と異なった現象が観察された。前者では，動詞の基となる形（例「思う」）に新たな要素が付加され，新しい形（例「思うから」）が使われること，後者では初出形に日本人幼児は普通体，学習者は丁寧体が多く使用されることがわかった。また，動詞の発達段階で，学習者特有の「動詞普通体＋です」（例「思ったです」）の中間言語形が出現し，「動詞普通体＋んです」（例「思ったんです」）の過渡的段階の形式であると推測された。

Abstract: In second language acquisition research, learners' language data are indispensable. In the sub-project "Research on the acquisition of Japanese as a second language in different learning environments," we developed a learners' language corpus called C-JAS. This report describes the nature of C-JAS and the acquisition patterns of Japanese verbs as observed in C-JAS. The unique feature of C-JAS is a longitudinal spoken corpus of three native speakers of Chinese and three native speakers of Korean, which has been tagged for morphological information and for learner errors, and which allows systematic searching. A C-JAS survey of the first appearances of the verbs *omou* 'think' and *taberu* 'eat' has revealed that there are phenomena both similar to and different from the patterns of first language acquisition. Deriving a new form (e.g. *omou kara* 'because (I) think') based on an earlier form (e.g. *omou* 'think') is common to both first and second language acquisition, whereas plain verb forms characterize Japanese children's verbs in their first appearances in contrast to the polite forms seen in learners' verbs. In the development of verb acquisition, a unique learners' interlanguage form "plain verb + *desu*" (e.g. *omotta desu* 'thought COP (POLITE)') appears, which seems to be a transitional form for "plain verb + *n desu*" (e.g. *omotta-n desu* 'it is the case that (I) thought so').

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，センター長。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），『日本語学習者による文法習得』（共著，くろしお出版，2001），『日本語教育に生かす第二言語習得研究』（アルク，2002），『プロフィシェンシーを育てる～真の日本語能力をめざして～』（共編著，凡人社，2008），『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）．

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，日本語教育学会国際連携委員会委員長，文化審議会専門委員（国語分科会）．

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」
サブプロジェクト「学習者の言語環境と日本語の習得過程に関する研究」
サブ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

本サブプロジェクトは，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に関する研究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，(1)「母語環境と第二言語環境」「教室指導環境と自然習得環境」などの学習者を取り巻く言語環境の違いが日本語習得に及ぼす影響に関する研究，(2) 学習者の母語が日本語習得に及ぼす影響（言語転移）に関する研究，そして，(3) そのための基礎資料として有用な日本語学習者の発話や作文のコーパスの内容と構造に関する研究を行う。