研究論文

日本語教育における学習者コーパスの構築とICLEAJ

Construction of Learner Corpora and International Corpus of Learner Japanese (ICLEAJ) in Japanese Language Education

望　月　通　子
MOCHIZUKI, Michiko

This study has three purposes. Firstly, I summarize the standards and characteristics found in currently available corpora of English among Japanese learners and of Japanese among foreign learners. Secondly, I discuss what aspects such learner corpuses might uncover through Granger's CIA model and what corpus designs would be able to address the criticisms against natural, or non-elicited, language data. Thirdly, I describe the design standards and characteristics found in our International Corpus of Learner Japanese（ICLEAJ）.

キーワード
学習者日本語コーパス　CIA　作文コーパス　コーパスデザイン　ICLEAJ

1．はじめに

1.1　本研究の背景

学習者コーパス研究が最も進んでいるのは英語教育であるが、1990年代初頭にはすでに研究者やEFL専門家、出版社が電子化学習者コーパスの理論的、実践的な可能性を認識し、プロジェクトを立ち上げた。学習者コーパスとしては、「ICLE」（International Corpus of Learner English、国際学習者英語コーパス、母語数16、サイズ200万語）、「LLC」（Longman Learners' Corpus、ロングマン学習者コーパス、母語数20、サイズ1,000万語）、「HKUST（Hong Kong University of Science and Technology）Corpus of Learner English」香港科技大学学習者英語コーパス、入学試験と定期試験の英作文、サイズ2,000万語）などがよく知られている。たとえば、「ICLE」は1990年に始まったプロジェクトで、2009年に初版のCD-ROM（ICLE V1）が、2009年にICLE V2が公開されたが、上級英語学習者（大学3～4年生）から一人500語以上の論説文を集めて編纂された、200万語規模の学習者英語コーパスである。筆者らが邦訳したGranger（1998）の*Learner English on Computer*は、同コーパスの構築や同コーパスに基づ

く研究成果を収録したものである。初版の CD-ROM（2003）には日本語話者の英作文モジュールはないが、第 2 版（2009）には日本人大学生 200 人分のデータがモジュールとして包含されている。

1.2　日本人の学習者英語コーパス

一方、日本国内における学習者英語コーパスについては、編纂・公開という形で結実したのは近年のことであり、2004 年～ 2008 年に相次いで一般公開されている。この他に、各機関や個人の私用に供するために編纂された、非公開の学習者英語コーパスも徐々に増えている。表 1 は、現在、一般公開されている学習者英語コーパスを一覧にしたものである。2004 年に「NICT JLE コーパス」（NICT Japanese Learner of English Corpus, NICT 日本人英語学習者コーパス）が公開されたが、アルクの SST（Standard Speaking Test）受験者 1,281 人から集めた 200 万語の英語会話データである。次いで、2007 年に「JEFLL コーパス」（Japanese EFL Learner Corpus, 日本人 EFL 学習者コーパス）が公開されたが、これは 10 年にわたって 1 万人の中高校生から集めた英作文コーパスである。翌年 2008 年には、大学生や大学院生の作文を集めた「NICE」（Nagoya Interlanguage Corpus of English、名古屋大学英語中間言語コーパス）と、アジア圏大学生の英作文を集めた「CEEAUS」（Corpus of English Essays Written by Asian University Students、アジア圏英語学習者コーパスネットワーク）が一般公開されている。

以上のように英語教育においては学習者英語コーパスの整備が急速に進展したが、本研究ではまず英語教育に追随する形で発展してきた日本語教育における学習者日本語コーパスの現状を概観し、次いで筆者らが構築した ICLEAJ の必要性を述べ、その構築の詳細について説明する。

2．学習者日本語コーパス

2.1　背景

英語はコーパス編纂やコーパス言語学が最も進んでいる言語である。1960 年代の電子コーパス第 1 号である「Brown コーパス」の登場以来、コーパスは辞書編纂から言語教育に至るまで様々な分野に浸透し、大規模化と多様化の一途をたどってきた。1994 年に完成した世界最大のイギリス英語コーパス「BNC」（British National Corpus）は大規模化の中で、一連の学習者英語コーパスの公開は多様化の中で生まれた産物といえるだろう。

2.2　汎用日本語コーパス

一方、日本語の場合はというと、その表記が多様で複雑であるうえに、非分かち書き言語であるということがネックとなり、コーパス整備が大幅に遅れたことは否めないだろう。国立国語研究所が「CSJ」（Corpus of Spontaneous Japanese、日本語話し言葉コーパス）を完成させ

表 1. 公開されている日本人英語学習者コーパス

文体	学習者コーパス名	属性および特徴
会話	NICT JLE コーパス	⑴編　纂：和泉絵美・伊佐原均・内元清貴（独立行政法人・情報通信研究機構）が構築したコーパス ⑵公開年：2004 年 ⑶サイズ：SST 受験者 1,281 人の 200 万語の会話データ ⑷タスク：SST ⑸習熟度：SST 判定結果があるので、習熟度別定量比較が可能である。 ⑹比較コーパス：サブコーパスが 2 つある。 「正解コーパス」：SST に準じるテストを受験した母語話者の音声データを書き起こしたサブコーパス 「日本語訳コーパス」：NICT JLE コーパスの一部を日本語に訳したサブコーパス
作文	JEFLL コーパス	⑴編　纂：投野由紀夫を中心に構築したコーパス ⑵公開年：2007 年 ⑶サイズ：10 年以上にわたって収集した約 1 万人の中高校生による 67 万語の英作文データ ⑷タスク：6 トピックから 1 つ選び、辞書なしで 20 分書く。 ⑸習熟度：なし、習熟度別データ比較はできないが、学年別定量比較や経年的定量比較は可能である。 ⑹比較コーパス：なし
作文	NICE	⑴編　纂：杉浦正利を中心に構築 したコーパス ⑵公開年：2008 年 ⑶サイズ：平均語数 337 語、ファイル数 207、総語数 69,858 語の大学生や大学院生の英作文データ ⑷タスク：11 のトピックから 1 つ選び、辞書なしで 60 分書く。 ⑸習熟度：TOEIC,TOEFL のスコアがあるので、習熟度別定量比較が可能 ⑹比較コーパス：平均語数 588 語、ファイル数 200、総語数 117,571 語の英語母語話者コーパスがある。
作文	CEEAUS	⑴編　纂：石川慎一郎を中心に構築したコーパス ⑵公開年：初版（2008）、第 2 版（2009） ⑶サイズ：「日本人英語学習者コーパス」（CEEJUS）はデータ数 770、総語数 169,654 語の大学生の英作文データ ⑷タスク：2 トピックから 1 つ選び、辞書なしで 20 ～ 40 分で 200 ～ 300 語書く。 ⑸習熟度：TOEIC（R）型推定スコアの 4 段階区分データがあるので、習熟度別定量比較が可能である。 ⑹比較コーパス：「日本人英語学習者コーパス」以外にモジュールが 3 つある。 「中国人英語学習者コーパス」（CEECUS） 「英語母語話者コーパス」（CEENAS） 「日本語母語話者コーパス」（CJEJUS）

たのは 2004 年のことであるが、雑誌「太陽」をデータ化した「太陽コーパス」が登場したのは翌 2005 年のことである。5 か年計画で BNC に匹敵する約 1 億語の大規模コーパス BCCWJ（Balanced Corpus of Contemporary Written Japanese, 現代日本語書き言葉均衡コーパス）を編纂し、公開したのは、2011 年 8 月のことである。

2.3　学習者日本語コーパス

英語教育に追随する形で日本語の学習者コーパスも発展しつつある。

下掲の表 2 は、現在、一般公開されていて、研究者や教育者の間で普及している日本語学習者コーパスを一覧表にまとめたものである。すでに述べたように英語学習者コーパスはその種類や容量の面で 2004 年～ 2008 年の間に一気に整備されたといえるだろうが、学習者日本語コーパスに関しては、2000 年前後から、日本語会話を集めた「KY コーパス」、日本語作文データを収集した「作文対訳 DB」が公開されていることがわかる。最初から学習者コーパスとしてデザイン・編纂された学習者コーパスとしては、台湾、英国、ウクライナの大学生の日本語作文を集めたオンラインの「日本語学習者言語コーパス」が 2009 年に、次いで、2011 年に 55 か国からの留学生の日本語作文を収集した「JLPTUFS」が一般公開された。「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）」には日本人と留学生の接触場面を含む様々な種類の会話を収集している。

以上のように 2009 年～ 2011 年の間に一気に整備が進んだといえるだろう。ここでは公開されているものだけを取り上げたが、各機関や個人の私用に供するために編纂されているものも増えている（名大日本語学習者コーパス他）。なお、学習者コーパスではないが、すでに 1990 年に報告書が公開されている『外国人学習者の日本語誤用例集』の PDF 版、DB 版が 2011 年に公開されているので、参考までに表に含めた。

SLA（Second Language Acquisition, 第二言語習得）は外国語や第 2 言語の学習プロセスを支配する原理を明らかにすることを主要目的にしているが、Ellis（1994: 670）はそのデータには学習者の言語使用データに加え、文法性判断のようなメタ言語的判断や質問紙法や思考発話法のような自己申告データなど 3 種類あるとしている。さらに、非統制の言語使用データを非誘導型、有統制のそれを誘導型、メタ言語的判断や自己申告データを内省型データと呼んで、優れた研究は様々なデータソースを使用している研究であると述べている。そしてこれまでの SLA 研究は、内省型や誘導型の言語使用データを中心に行われてきたが、学習者コーパスは非誘導型の言語使用データがもっている出現頻度や変数、回避といった問題を解決できる利点を十分に備えていると述べている。

次章では改めて「学習者コーパスとは何か」「それで何ができるのか」について考えてみたい。

3．学習者コーパスとは何か、それで何ができるのか？

3.1　背景

日本語学習者コーパスは、JFL 環境下あるいは日本での JSL 環境下で学習している日本語学習者から日本語の会話や作文を収集し、それと並行して日本語話者の母語としての日本語の会話や作文を収集することで、外国人話者と日本人話者間の日本語の比較検討が可能になる。日本人話者の日本語の会話や作文のデータを参照コーパスとして外国人学習者の過剰使用や過少使用の傾向が明らかになり、L1 転移や日本語学習者の回避方略、母語話者あるいは非母語話者

表 2. 公開されている外国人日本語学習者コーパス

文体	学習者コーパス名	属性および特徴
会話	KY コーパス	⑴編　纂：鎌田修・山内博之が主体となり構築したコーパス ⑵公開年：初版（1999）、修正版 1.2 版（2004） ⑶サイズ：日本語の OPI で受験者が発話した発話を文字化した 90 人分のデータ。日本語学習者の母語はそれぞれ中国語、英語、韓国語で 30 人ずつある。 ⑷タスク：OPI（口頭能力試験） ⑸習熟度：評価レベルは初級 5 人、中級 10 人、上級 10 人、超級 5 人。習熟度別定量比較が可能である。 ⑹比較コーパス：なし
会話	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）	⑴編　纂：宇佐美まゆみを中心に構築したコーパス ⑵公開年：2009 年版 , 2011 年版（増補版） ⑶サイズ：294 会話、総時間 4000 分 31 秒（約 66 時間）の会話が収録されており、そのうち音声付きデータは 136 会話、1164 分 43 秒（約 20 時間） ⑷タスク：様々な状況の会話 ⑸習熟度：各会話グループの実験計画や話者の年齢・性別・属性等のデータベースがある。 ⑹比較コーパス：日本語母語話者同士、日本語母語話者と日本語学習者の会話が含まれている。
作文	『外国人学習者の日本語誤用例集』	⑴編　纂：寺村秀夫を中心に収集 ⑵公開年：「外国人学習者の日本語誤用例の収集・整理と分析」の資料をまとめた報告書 1990、PDF 版／データベース版 2011 ⑶サイズ：20 か国、延べ 339 人の日本語作文、420KB ⑷タスク：自由作文、パターン作文、短文作文、聴解要約、会話作文、絵を見ての作文 ⑸習熟度：表示なし　国籍、作文形式、誤用の種類による定量分析は可能 ⑹比較コーパス：なし
作文	作文対訳 DB	⑴編　纂：宇佐美洋（国立国語研究所）を中心に構築した DB ⑵公開年：初版（2000）、増補版（2001）、再増補版（2009） ⑶サイズ：2009 年版は、20 か国の学習者作文と日本語母語話者作文が合計 1,500 編収集されている ⑷タスク：300 ～ 800 字程度の日本語作文 ⑸習熟度：習熟度情報がないため習熟度別定量比較はできないが、執筆者・添削者の言語歴情報があるので、学習期間別定量分析は可能 ⑹比較コーパス：執筆者本人による作文の母語訳、学習者作文の添削、日本語母語話者作文
作文	日本語学習者言語コーパス	コーパス（日本語誤用オンライン辞書も公開） ⑴編　纂：海野多枝（東京外国語大）を中心に構築したコーパス ⑵公開年：2009 年版、2010 年版、2011 年版 ⑶サイズ：2011 年版の作文数は 1,756 編、総語数は 267442 語 ⑷タスク：作文や日記タスク、機能タスク ⑸習熟度：習熟度表示なし、 ⑹比較コーパス：台湾日本語学習者データ、英国日本語学習者データ、ウクライナ日本語学習者データ、日本語母語話者データなどがあるので、母語が異なる学習者間、学習者と母語話者間の定量比較が可能
作文	JLPTUFS 作文コーパス	⑴編　纂：東京外大留学生別科の教員を中心に構築したコーパス ⑵公開年：2011 ⑶サイズ：入門～超級まで 8 レベルの 55 か国 1,515 編 ⑷タスク：授業や自宅学習の作文 ⑸習熟度：日本留学試験や日本語能力試験の得点の情報はないが、クラスや国籍の情報があるので、クラス別、1 ～ 2 級別、国籍別の定量分析は可能 ⑹比較コーパス：なし

に特徴的な言語使用特性、ならびに日本語学習者の苦手とする言語領域など、そこに見られる発見が日本語教育の教材作成に資することにもなる。

3.2　CIA（Contrastive Interlanguage Analysis 対照中間言語分析）

1960 年代を通じて CA（Contrastive Analysis　対照分析）が一世を風靡し、母語と外国語の類似点と相違点を比較して習得の難易度や誤りを予測することが重視された。しかし、学習者のエラーを分析した結果、対照分析が予測するエラーと実際に生じるエラーに矛盾が見られ、2 言語間の違いに基づいて習得の難易度を決定することにも問題があることがわかり批判にさらされることになった。

次に登場したのが Corder（1967）の EA（Error Analysis 誤用分析）である。エラーは TL（Target Language 目標言語）の習得過程で学習者自身がたてた仮説を検証する中で生じるもので、不完全ながらも体系性を備えているとした。Selinker（1972）はこの不完全な TL に到達するまでの段階の言語体系を IL（Interlanguage 中間言語）と呼んだ。しかし、エラーがないことと学習者による回避との区別がつかないこと、データのサイズが小規模であること、習得に影響する変数の特定が難しいことなどにより、新たな批判にさらされることになった。

第 1 章で述べたとおり、1990 年代初頭に学習者コーパスのプロジェクトがスタートしているが、学習者コーパス研究はコーパス言語学と SLA 研究に根ざし、コーパス言語学の手法を使ってオーセンティックな学習者言語をもっと深く洞察しようとするものである。（Granger 1998、翻訳 2007 xix）

コーパス言語学が開発した手法の活用により、以前の CA に欠けていた科学的経験的なアプローチが可能になり、Granger（1996）はこれを CIA（Contrastive Interlanguage Analysis）と呼んでいる。Odlin（1989:212, Granger 1996:43）は、このような比較は転移の考察に有益であるとしている。図 1 は ICLEAJ の場合についての CIA を示したもので、次の 2 種類を比較する（Granger：1996）。

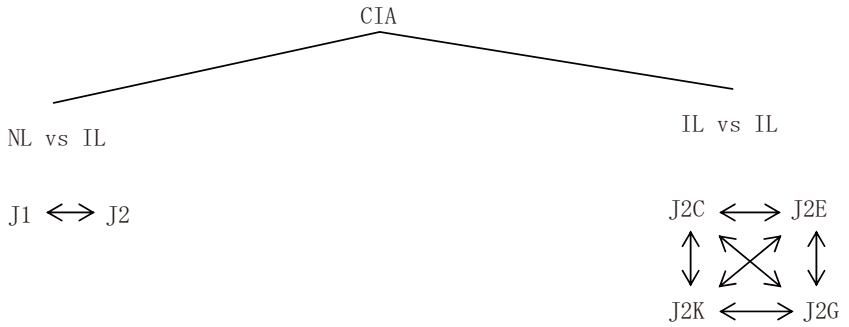

図 1　Contrastive Interlanguage Analysis（Granger1966：44 Figure に基づき、筆者が修正）

⑴ NL 対 IL　同一言語の母語と非母語の変種間の比較。例えば、母語話者の日本語（J1）と外国語としての日本語（J2）を比較する。

⑵ IL 対 IL　同一言語の様々な中間言語間の比較。例えば、中国人学習者（J2C）、韓国人学習者（J2K）、米国人学習者（J2E）、ドイツ人学習者（J2G）間の日本語を比較する。

4.「国際学習者日本語コーパス」の編纂

以下、KCOLJ（関大学習者日本語コーパス）および KCONJ（関大母語話者日本語コーパス）を拡充した「ICLEAJ」（International Corpus of LEArners of Japanese、国際日本語学習者コーパス）の特徴とその構築について説明していく。

4.1　「ICLEAJ」の必要性

第 2 章で公開されている外国人日本語学習者コーパスを概観したが、作文コーパスには「作文対訳 DB」「日本語学習者言語コーパス」「JLPTUFS 作文コーパス」がある。いずれも母語別のモジュールがあり異言語間（NL vs. IL, IL vs. IL）の比較は可能であるが、JFL vs. JSL、作文の種類による比較には十分とはいえない。

4.2　「ICLEAJ」の基本設計

上掲の改善点を考慮して、ICLEAJ では以下の点について統制を行っている。

学習環境：JSL 環境／ JFL 環境を明記

母　語：母語を明記

学習者：外国語または第二言語として日本語を学習している学生・大学院生に限定

習熟度：日本語能力試験／ N 試験のレベルを明記（記載がある場合）

タスク：20 種類のテーマを設定している。論述文（意見文）10 テーマ、叙述文（経験文）10 テーマより 1 つ選び、テーマは以下のとおりである。1 時間で 800 字程度の作文を手書きまたはワードで作成させる。なお、作文時の辞書使用は許可している。

4.3　「ICLEAJ」の概要

「ICLEAJ」は、「ICLEAJ-NNS」と比較用の統制コーパス「ICLEAJ-NS」から成る。前者の「ICLEAJ-NNS」は、JSL 環境下の学習者データと JFL 環境下のそれを区別している。CN（中国語話者）の作文は、「NNS-JSL-CN」（中国語話者 JSL 200 編）と「NNS-JFL-CN」（中国語話者 JSL 200 編）のモジュールがある。KR（韓国語話者）の作文は JL 環境の「NNS-JSL-KR」（韓国語話者 JSL 100 編）と「NNS-JFL-KR」（韓国語話者 JFL 100 編）で構成されている。後者の「ICLEAJ-NS」は、大学生・大学院生に限定した「NS-JPU」（200 編）と一般社会人による

「NS-JP」（200 編）で構成され、中年層と高年層に分けることができる。

4.4　データの電子化と ICLEAJ のβ版の公開

収集されたデータは、コーパスデータとして処理可能な状態にするため、テキストファイル化して SJIS により保存した。学習者コーパス 79 名分（JSL 中国語話者 43 名分、JFL 韓国語話者 36 名分）および日本語母語話者コーパス 86 名分（日本人学生作文 44 名分、日本人社会人作文 42 名分）のデータをβ版として 8 月 19 日から公開している。データの文字コードは EUC-JP、改行コードは LF（UNIX 系）となっている。

5　今後の課題

ICLEAJ のデータは、エラータグおよびモデル文を付加したうえで、2013 年 2 月に公開する予定である。また、JSL, JFL 各環境の母語別サブコーパス構築への協力者を募集し、コーパスデータを拡充する予定である。

謝辞　本研究は科学研究費による研究の一環として行ったものであり、ここに感謝を記したい。課題番号 22520543（研究代表者：望月通子、分担研究者：阪上辰也）

参考文献

Corder, S. P.（1967）. *The Significance of Learner's Errors. An Introduction.* Oxford:Basil Blackwell.

Ellis, R.（1994）. *The Study of Second Language Acquisition*, Oxford: Oxford University Press.

Ellis, R. & Barkhuizen G.（2005）. *Analyzing Learner Language.* Oxford: Oxford University Press.

Gilquin, G., Rapp, S., & Diez-Bedmar, M.B.（Eds.）.（2008）. *Linking up Contrastive and Learner Corpus Research*. Amsterdam, The Netherlands: rodopi.

Granger, S.（1996）. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora.In K.Aijmer, B. Altenberg, & M. Johansson（Eds.）, *Language in Contrast: Text-based Cross-Linguistic Studies*（pp.37-51）. Lund, Sweden: Lund University Press.

Granger, S.（1998）. *Learner English on Computer.* Harlow, England:Addison Wesley Longman［船城道雄・望月通子（訳）、（2008）『英語学習者コーパス入門：SLA とコーパス言語学の出会い』、東京：研究社］

Granger, S., Dagneaux, E., & Meunier, F.（Eds.）.（2003）. *International Corpus of Learner English*. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M.（Eds.）.（2009）.*International Corpus of Learner English*. Version 2. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

石川慎一郎（2012）『ベーシックコーパス言語学』東京：ひつじ書房

和泉絵美・内元清貴・井佐原均（編）（2004）『日本人 1200 人の英語スピーキングコーパス』東京：アルク

Leech, G.（1998）. Preface. In S.Granger（Eds.）, *Learner English on Computer*（pp.xiv-xx ）. Harlow, England: Addison Wesley Longman.

前川喜久雄（2011）「特定領域研究『日本語コーパス』と『現代日本語書き言葉均衡コーパス』」『「現代日本語書き言葉均衡コーパス」完成記念講演会予稿集』、1–10

Odlin, T.（1989）. *Language Transfer Cross-linguistic Influence in Language Learning*. Cambridge: Cambridge University Press.

Selinker, L.（1969）. Interlanguage. *International Review of Applied Linguistics in Language Teaching* 10: 209–231.

阪上辰也・杉浦正利・成田真澄（2008）「学習者コーパス『NICE』の構築」杉浦正利（編）『平成 17 ～ 19 年度科学研究費補助金基盤研究（B）研究成果報告書：英語学習者のコロケーション知識に関する基礎的研究』（pp.1–14）名古屋」名古屋大学

宇佐美洋（2001）『平成 11 ～ 12 年度科学研究費補助金基礎研究（B）（2）研究成果報告書：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』東京：国立国語研究所

宇佐美洋（2002）「『対訳作文データベース』と日本語教育：対照言語学を教育に生かすために」国立国語研究所（編）『日本語と外国語との対照研究 X：対照研究と日本語教育』（pp.82–94）。東京：国立国語研究所

海野多枝・鈴木綾乃（2011）「中級日本語学習者コーパスに見られる語彙的コロケーション：動詞『する』を中心に」『コーパスに基づく言語学教育研究報告』（東京外国語大学）、7、327–345

鎌田修（2006）「KY コーパスと日本語教育研究」『日本語教育』、*130*、42–51

寺村秀夫（1990）『外国人学習者の日本語誤用例集』（大阪大学；PDF 版、国立国語研究所、2011 年）

寺村秀夫（1990）『外国人学習者の日本語誤用例集』（大阪大学；データベース版、国立国語研究所、2011 年）

投野由紀夫（編）（2007）『日本人中高生一万人の英語コーパス：中高生が書く英文の実態とその分析』東京：小学館

山内博之（n.d.）「KY コーパス」日本 OPI 研究会ウエブサイト「OPI を利用したコーパス」Retrieved from http://opi.jp/shiryo/ky_corp.html（2012.5.20）