東京外国語大学
留学生日本語教育センター論集　36：123〜133，2010

JLPTUFS 作文コーパスの構築について
—全学日本語プログラムで学ぶ日本語学習者の作文データベース化—

鈴木 智美・中村　彰・韓　金柱[1]
（2009. 10. 31 受）

【キーワード】 作文コーパス、全学日本語プログラム、テキストファイル、
教育 GP「世界的基準となる日本語スタンダーズの構築」

1　はじめに

本稿では、留学生日本語教育センター（以下、留日センター）において進められている「JLPTUFS 作文コーパス」構築プロジェクト（2008 〜 2010 年度）について、2009 年度秋現在の中間報告を行う。

2　プロジェクトの概要

プロジェクトの概要は、以下の通りである。

(1)「JLPTUFS 作文コーパス」プロジェクト

内容：東京外国語大学「全学日本語プログラム」（JLPTUFS）[2] の教育課程において書かれた作文のうち、執筆者によるデータ提供の同意が得られた作文をデータ化し、日本語学習者の作文コーパスを構築する。

位置付け：留日センター「教育研究開発プロジェクト」

---

[1] 韓金柱は東京外国語大学大学院地域文化（現総合国際学）研究科博士後期課程在籍。本プロジェクトの教務補佐としてデータベース化作業の主要な部分を補佐している。

[2]「全学日本語プログラム」（JLPTUFS:Japanese Language Program, Tokyo University of Foreign Studies）は、2004 年（平成 16 年）4 月より全学向けに開かれることとなった日本語教育プログラムである。留日センターに運営委員会をおき運営が行われている。交流協定校からの交換留学生（ISEP-TUFS（International Student Exchange Program, TUFS）等で学ぶ短期留学生）、日本語・日本文化研修留学生、教員研修留学生、国費および私費の研究生など、全学の非正規生の留学生（一部正規の大学院生を含む）を対象とする。100（入門）レベルから 800（超級）レベルまで 8 段階のレベルがあり、2009 年度秋学期現在、計 44 に及ぶ各種の日本語科目が開講されており、1 週間の延べ開講コマ数は、計 89 に上る（1 コマ＝90 分の授業）。初級レベルでは 1 週間に 10 コマの授業が行われる集中コースも設けられており、国費研究留学生の予備教育にも対応している。

期間：2008年度～2010年度

予算：文部科学省「質の高い大学教育推進プログラム」（教育GP）[3] 平成20年度選定取組「世界的基準となる日本語スタンダーズの構築」（東京外国語大学）による（2008年度秋より）

担当者[4]：鈴木智美、中村彰

(2) 目的

日本語学習者の作文を電子データ化し、数多く蓄積していくことで、作文における文法項目、語彙、漢字等の用いられ方、および学習している人の母語や学習レベルと作文との間にどのような関係があるか等を分析し、日本語教育に生かすことを目的とする。

(3) 本コーパス作成の価値および利点

- 母語背景、日本語レベル等、多様な学習者の作文データが収集できる。
- 横断的、かつ縦断的（学習期間1年の）データ収集ができる。
- シンプルなテキスト形式でデータベースを構築することにより、誤用研究・コーパス研究等に役立つ基礎的な材料を提供することができる。
- 公開可能なコーパスを構築することで、学内・学外を問わず日本語教育界全体に貢献することができる。

(4) 作文収集期間および対象クラス

収集期間：2009年度～2010年度（春・秋学期 ×2年間＝計4学期分）

---

[3] 文部科学省「質の高い大学教育推進プログラム」（Program for Promoting High-Quality University Education）は、教育の質の向上につながる特に優れた教育取組に対し、社会への情報提供を行うとともに、重点的な財政支援を行い、国全体としての高等教育の質の保証、国際競争力の強化に資することを目的として設置されたとされるものである。平成20年度には148件の取組が選定されている。（文部科学省 http://www.mext.go.jp/a_menu/koutou/kaikaku/gp/program/08033118.htm 参照）

[4] 2008年度は伊集院郁子氏を加え、3名体制でプロジェクトを進めた。2009年度からはデータ収集の対象となる日本語教育プログラムによってプロジェクトを2つに分けることとし、本プロジェクトの担当者は鈴木智美、中村彰の2名となった。留日センター国費学部留学生予備教育プログラム（「1年コース」）の教育課程における作文についても、2009年度より、別途伊集院氏を中心にデータベース化のプロジェクトが進められている。

対象クラス：「集中」[5] コース、「総合」[6] クラス、「文章表現」[7] クラス、
「アカデミック・ライティング」[8] クラス

本プロジェクトは、「全学日本語プログラム」の教育カリキュラムの充実化を図る中で、2007年度秋に構想がスタートした[9]。2008年度には、留日センターにおける教育研究開発プロジェクトの1つとしてプロジェクトが発足し[10]、同年10月より「質の高い大学教育推進プログラム」（教育GP）「世界的基準となる日本語スタンダーズの構築」の予算配分を受けて、計画が進められることとなった[11]。

---

[5] 100（入門～初級）レベル、および200（初級後半～初中級）レベルは、90分の授業が週に10回（毎日2つずつ）行われる「集中」コースの形態をとっている。

[6] 300（初中級）レベルから600（上級前半）レベルまでは、日本語の力を総合的にバランスよく身につけることを目標とする「総合」クラスと、「読解」「聴解」「口頭表現」「文章表現」等の各種「技能別」クラスから構成される。「総合」クラスは、90分の授業が週5回（毎日1つずつ）、技能別クラスはいずれも週1回行われる。700（上級後半）レベルも同様の構成だが、「総合」クラスは週に2回となる。

[7] 300（初中級）レベルから700（上級後半）レベルで開講されている「文章表現」のクラスである。週に1回の授業が行われる。

[8] 800（超級）レベルは、「文学日本語」「ビジネス日本語」「ドラマ・ドキュメンタリー」などのテーマ別の構成になっている。授業はいずれも週に1回行われる。文章表現に相当するクラスとして「アカデミック・ライティング」のクラスが設定されている。

[9] 作文コーパスの作成そのものについては、2007年6月には留日センターにおける日本語教員全体会議にて本稿の執筆者の一人（鈴木智美）より呼びかけがなされた。この計画は、具体的には「英語力・日本語力高度化推進プロジェクト」（2008年度特別教育研究経費に応募）の一環として、2007年秋より現実的にスタートすることとなった。なお「英語力・日本語力高度化推進プロジェクト」のうち、留日センターが担当する日本語力高度化（「全学日本語プログラム」の充実）の計画部分については、2008年度春学期は学内予算（学長裁量経費）により実施することとなった。

[10] 留日センターでは、各教育コース・プログラムの運営のほかに、教育研究開発に関わる各種プロジェクトをセンター内公募の形により毎年複数実施している。「全学日本語プログラム」を対象とした作文コーパスの作成については、2008年度に新規プロジェクトの1つとして立ち上がった（担当者：鈴木智美、中村彰、伊集院郁子）。プロジェクト予算は留日センター教育改革費（留日センター長裁量経費）であったが、この作文コーパス作成は、具体的には2008年度秋学期より始動するプロジェクト計画であったため、2008年度には、このプロジェクトによる留日センター上記経費の使用は生じていない。

[11] 2008年度秋に教育GP予算の配分を受け、作文コーパス作成の対象を「全学日本語プログラム」だけでなく、国費学部留学生予備教育プログラム（「1年コース」）にも広げようという方向性が「スタンダーズGP協議会」（留日センターにおいて上記GPを円滑に進めていくために設置された会）より生まれた。ただし、両者のコーパスは基本的な設計図および公開可能性について同一に扱うことが難しいため、2009年度よりプロジェクトを2本に分割して進めていくこととした。

3 「JLPTUFS 作文コーパス」の設計図

3.1 イメージサンプル

「JLPTUFS 作文コーパス」は、以下の 3 つのデータから構成される。

a. 情報一覧ファイル（EXCEL 形式）
b. 作文テキストファイル
c. 作文 PDF ファイル

以下の図 1 に示すのは、そのイメージサンプルである。ただし、以下はサンプルであり、実際には、情報一覧ファイルの項目記載順などは、この通りではない。

情報一覧ファイルの作文番号から、当該作文のテキストファイルおよび PDF ファイルへとリンク付けがなされ、コーパス使用者は当該作文のファイルを見ることができる。執筆者氏名や、作文中の個人名は削除した上でデータ化を行うこととする。

a. 情報一覧ファイル（EXCEL 形式）

作文番号	PDFファイル	執筆者ID	レベル	クラス	国籍	母語	専門	作文のテーマ	作文タイトル	実施形態	制限時間	字数指定
20091100100051901	20091100100051901	10005	100	集中	フィリピン	フィリピン語	国際関係	わたしの1日	わたしの一日	宿題	特になし	約400字
20091100100062307	20091100100062307	10016	100	集中	アイルランド	英語	美術	日本に来る前と日本に来てから	日本のせいかつ	宿題	特になし	500-600字
20091200100072106	20091200100072106	10010	200	集中	ブラジル	ポルトガル語	史学	留学と将来の夢	わたしの将来のゆめ	授業時間内	90分	約600字
20091300300052110	20091300300052110	10024	300	文章	中国	中国語	教育学	自国と日本の習慣の違い	中国と日本の習慣の違い	宿題	特になし	約400字
20091400200071305	20091400200071305	10013	400	総合	アメリカ合衆国	英語	文化人類学	自国の教育制度	アメリカペンシルベニア州の教育制度について	宿題	特になし	特になし
20091700200050814	20091700200050814	10135	700	総合	ウズベキスタン	ウズベク語	言語学	「機械との共存」を読んで考えたことについて書きなさい	人間の未来	授業時間内	45分	800字以内
20091800400616O1	20091800400616O1	10120	800	アカデミックライティング	ロシア	ロシア語	日本語・日本文化	因果関係を述べる	ロシアにおける頭脳流出	宿題	特になし	400字

sample

b. 作文テキストファイル

（国籍：フィリピン）
（母語：フィリピン語）
（テーマ：わたしの1日）
（入力日：090908）
（入力者：GO）
（確認日：091020）
（確認者：KK）
（注：一部ルビあり）

わたしの一日
　わたしは、毎朝５時半ごろおきます。朝いちばんにすることはおいのりです。それから、おちゃをのみます。そして、日本ごをべんきょうします。朝ごはんをたべてから、シャワーをあびます。７時１５分ごろりょうをでます。
　わたしは仙川でけいおうせんにのって、とびたきゅうえきでおりて、大学まであるいて行きます。そしがやから大学まで１時かん半ぐらいかかります。
　わたしのじゅぎょうは９時にはじまります。日本ごセンターの３０８ごうしつで、日本ごをべんきょうします。そのきょうしつはあたらしくて、あかるいです。先生のこうぎの前に、わたしは本をよみます。よくふくしゅうをします。日本ごのべんきょうはすこしむずかしいですが、とてもおもしろいです。この日本ごべんきょうはたいせつですから、わたしは毎日ならいます。

c. 作文 PDF ファイル

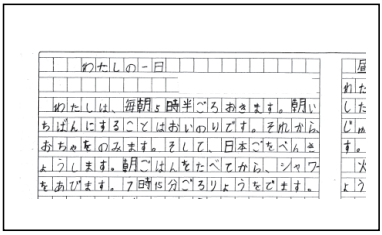

図 1 「JLPTUFS 作文コーパス」イメージサンプル

3.2　情報一覧ファイルの記載項目

情報一覧ファイル（EXCEL 形式）には、以下の項目が記載される。

(1) 基本情報：
　作文番号[12]、執筆者 ID 番号[13]、レベル[14]、クラス[15]

(2) 執筆者情報：
　性別、年齢、専門、国籍、国籍以外の 3 年以上の居住地、
　母語、母語以外に（国あるいは上記居住地で）日常的に使用していた言語、
　日本語能力試験の合格級および合格年（合格者のみ）

(3) 作文情報：
　実施日、出題テーマ、タイトル、実施形態（授業時間内実施、宿題など）、
　制限時間、字数指定、文体指定、筆記形態（手書き、ワープロの別）、
　その他条件（辞書使用可、教科書・ノート参照可など）

(1) の基本情報は、学期ごとに「全学日本語プログラム」運営委員会より受講者の基礎データの提供を受け、これを作文執筆者からのデータ提供同意書に照らして、作成する。(2) の執筆者情報は、データ提供に同意した執筆者より書面で情報を収集する。(3) の作文情報は、「全学日本語プログラム」の各クラス・コース担当教員の協力を仰ぎ、書面で収集する。作文の筆記形態については、個別に当該作文を確認する。

4　2008 年度の活動：設計・準備段階

2008 年度は、コーパスの設計・準備段階であり、以下の各点について進めた。

---

[12] 個々のデータには作文番号が付される。データの収集年度、学期、レベル、クラス、収集日などがわかるようにした 16 桁の番号となる。
[13] 同一の執筆者による作文は、執筆者 ID 番号により検索が可能となる。ただし、執筆者個人を特定する情報（氏名、学籍番号、留学カテゴリー等）については、一切非公開である。
[14] 「全学日本語プログラム」のレベル（100 ～ 800 レベル）を示す。
[15] 「全学日本語プログラム」のコース・クラス形態（「集中」コース、「総合」クラス、「文章表現」クラス、「アカデミック・ライティング」クラス）を示す。

(1) コーパス設計図の作成、作成手順の検討、およびサンプルの作成[16]
(2) 作文収集のためのデータ提供依頼書・同意書の作成[17]
(3) 必要機器の購入[18]
(4) 講演会および研究会の企画・開催[19](以下 a. ～ c.)

a. 教育 GP「世界的基準となる日本語スタンダーズの構築」講演会
タイトル:「学習者コーパスは役に立つか」
講演者:宇佐美洋氏
(独立行政法人国立国語研究所日本語教育基盤情報センター)
日時:2009 年 1 月 29 日(木)16:00 ～ 18:00

b. 作文コーパス研究会(1)
タイトル:「作文コーパスに望むこと－現場の日本語教員の目から考える－」
講師:家田章子氏(桜美林大学)
日時:2009 年 2 月 19 日(木)16:30 ～ 18:00

c. 作文コーパス研究会(2)
タイトル:「教育分野における Web デザイン実践のプロセスについて」
講師:角南北斗氏(ウェブディレクター・デザイナー)
日時:2009 年 2 月 23 日(月)16:00 ～ 17:30

---

[16] コーパスの設計図および作成手順については、鈴木智美が原案を作成し、2008 年度プロジェクトメンバーの中村彰、および伊集院郁子氏の協力を得て練り上げを行った。イメージサンプルは 5 回にわたって作成し、検討した。また、イメージサンプルの作成段階では、コーパス研究会の講師としても招聘した桜美林大学の家田章子氏にも、多くの助言をいただいている。

[17] 作文収集のためのデータ提供依頼書・同意書については、伊集院郁子氏が原案を作成し、同上、プロジェクトメンバーの 3 名で練り上げを行い、作成した。英語版の作成は中村彰が行った。

[18] 基本的なハード類、ソフト類の選定については、中村彰および伊集院郁子氏が主として担当した。

[19] 直接的には本プロジェクトにおける研究会ではないが、2007 年 11 月には、本プロジェクトメンバーの一人(鈴木智美)の発案で、留日センター FD(Faculty Development)研修会にて、「日本語研究とコーパス:日本語教育への応用を視野に」の演題で滝沢直宏氏(名古屋大学大学院国際開発研究科)にも講演を行っていただいている。

5　2009年度の活動：作文収集とデータベース構築

5.1　作文データの収集

2009年度からは、実際に作文の収集とデータベース化を進めている。作文の収集は、これまで以下の2回にわたり行っている。

(1) 作文データ収集（第1回）：2009年度春学期

- 全学日本語プログラム講師会での説明：2009年3月12日（木）
- 各クラスでの説明および同意書の収集[20]：2009年4月20日（月）～23日（木）

(2) 作文データ収集（第2回）：2009年度秋学期（継続中）

- 全学日本語プログラム講師会での説明：2009年9月17日（木）
- 各クラスでの説明および同意書の収集：2009年10月26日（月）～29日（木）

この結果、2009年秋現在、第1回の収集期間においては、表1（報告末尾に掲載）に示す数の作文データが収集された[21]。

作文の収集にあたっては、対象となる各クラス・コースの担当教員の協力を得て、実施作文のコピーおよび作文実施情報を提出していただいている。提出された作文コピーの中から、データ提供の同意を得ている執筆者によるもののみを選別した上で、データ化を行っていく。

5.2　データベース構築

収集された作文に基づき、データベース化の作業を以下のように進めている。

---

[20] 対象となる各クラスでの説明およびデータ提供の同意者募集については、鈴木智美および中村彰が分担して各クラスで行っている。時間割の都合上、両名が説明に回れない場合には当該クラス担当教員に説明および同意書回収をお願いすることとしている。「全学日本語プログラム」では、授業開始後2週間目に履修登録の締切日が設定されている。よって、各クラスの受講メンバーが落ち着いた頃として、授業開始後2～3週間目あたりにコーパスの説明日程を設けている。

[21] 表1に示すのはコーパス作成にあたって最終的に収集できた有効データ数であり、各クラス・コースで実際に行っている文章表現タスクの総数は、この収集データ数を上回るものである。コーパス構築にあたっては、複数回の書き直しを行った同一作文については、初稿のみをデータとして収集することとし、また、図表の説明を中心としたものや、文献からの引用を伴う長文の論文形式のものなど、コーパスとしてテキストデータ化の難しいものについては対象から除くこととしている。

(1) データベース化作業手順の策定、および作業マニュアルの作成[22]

(2) 情報一覧ファイルの作成[23]

(3) 作文PDFファイルの作成[24]

(4) 作文テキスト入力および確認作業手順の策定[25]

(5) 作文テキスト入力マニュアルの作成[26]

(6) 作文テキストファイルの作成[27]

(7) 作文テキストファイルの入力確認[28]

(8) 情報一覧ファイルと、各テキストファイル、PDFファイルの統合リンク付け、および全体の確認

---

[22] データベース化全体の作業手順については、鈴木智美が策定・指示し、韓金柱（作業補佐）がその手順を文書化して、マニュアル作成を行った。データベース構築についての作業手順を細かく記しておくこととしている。

[23] 3.2節に示したように、「全学日本語プログラム」の受講者データ、作文データ提供の同意書、作文の実施情報などを総合して、情報一覧ファイルを作成することになる。鈴木智美が指示し、韓金柱（作業補佐）が作成作業にあたっている。

[24] データ提供の同意書を得ている全作文にプロジェクト担当教員（2009年度春学期は鈴木智美、秋学期より鈴木智美および中村彰）が目を通し、指示された作文テーマに合致していないものや、字数指定との相違（字数の不足）が著しいもの、コピーの不鮮明なもの等はデータ化対象から除外することにしている。また、執筆者氏名や作文中の個人名など、削除すべき箇所を指示し、以上の指示を受けた上で、韓金柱（作業補佐）が全作文のPDF化作業を担当している。

[25] 大量の作文データを扱い、複数名の作業補佐が分担して作業にあたることになるため、全体の作業手順を細かく定めた上で、作業補佐と複数回の打ち合わせを行っている。プロジェクト担当教員（鈴木智美）が手順の策定と指示を担当している。

[26] テキストファイルの入力方法は、「JLPTUFS作文コーパス」と、「1年コース」（国費学部留学生予備教育プログラム）対象の作文コーパスとで、共通の方式をとることとし、細かくルールを決め、マニュアルを作成している。入力マニュアルの作成にあたっては、鈴木智美が原案を作成し、伊集院郁子氏、中村彰両名の協力を得て、第1版を作成した。入力マニュアルは、テキスト入力作業と平行して、随時改訂を重ねていくこととしている。改訂は鈴木智美が担当している。

[27] 5名の作業補佐にテキストファイル作成作業にあたってもらっている。レベル別に担当を分け、入力マニュアルに従い、まず1時間に作文1～2本のペースを目安に入力作業を進めている。入力にあたって生じた問題点・疑問点については、連絡ノートに記し、担当教員（2009年度春学期は鈴木智美、秋学期より鈴木智美および中村彰）がその解決を指示・記入し、必要に応じて入力マニュアルの改訂に反映させることとしている。各作業補佐は、毎回この連絡ノートを確認した上で作業に入ることとしている。なお、2009年度、入力作業を担当している作業補佐は、上久保明子、高橋希美、モンコンチャイ・アッカラチャイ、徐承希、左嵜遥香の5名である。

[28] 入力された全ファイルについて、2名の作業補佐に、2段階に分けて入力確認の作業にあたってもらうこととしている。マニュアル通りの入力が行われているかどうかをチェックする役割を担う。2009年度、この作業には、作業補佐として黄慧および上久保明子に担当してもらっている。

6　今後の課題

2009年度秋学期に収集された作文についても、上記5.2節と同様の手順でデータベース化を進めていく。また、2010年度にも同様の手順でデータベース化を行う予定である。

また、2010年度には、以下の各点について解決あるいは提案を行うことを考えたい。

(1)「全学日本語プログラム」の各レベルについてのレベル記述の改善
(2)「JLPTUFS作文コーパス」の実際の利用例案の収集
(3) 文章表現以外の学習者の産出データのアーカイブ化の検討

(1)については、作文コーパスに収められているデータについて、作文執筆者の日本語レベルを、対外的にわかりやすく対応付けることが必要だと考えられるためである。(2)については、2008年度の設計・準備段階から計画に入れたいと考えていた点である。コーパスの全体像が見えてきたところで、その実際の利用例についてまとめたいと考えている。(3)については、教育実践の成果として、またそれをこれからの日本語教育に役立たせるためにも、文章表現以外に、広く学習者の産出データのアーカイブ化を進めていくことには意義があるのではないかと考える。また、より視野を広げれば、学習者の産出データのみならず、日本語教授法・日本語教育の実践そのもののアーカイブ化についても、今後検討していく余地があるのではないだろうか。

表1　2009年度春学期の収集データ数

レベル	クラス	データ数	レベル別計
100	集中	22	22
200	集中	35	35
300	総合	28	
	文章	129	157
400	総合	25	
	文章	42	67
500	総合	43	
	文章	53	96
600	総合	0	
	文章	3	3
700	総合	15	
	文章	31	46
800	ｱｶﾃﾞﾐｯｸ･ﾗｲﾃｨﾝｸﾞ	72	72
総計			498

参考文献・資料

家田章子（2007）「語学教育におけるオープンソースソフト活用の可能性－MY Serverを利用して－」『第5回日本語教育研究集会予稿集』（於名古屋大学大学院国際言語文化研究科）pp.42-45

宇佐美洋（2009）「『学習者コーパス』は役に立つか？」東京外国語大学留学生日本語教育センター教育GP「世界的基準となる日本語スタンダーズの構築」講演会におけるハンドアウト（2009年1月29日）

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～平成10年度科学研究費補助金（基盤研究（A）（1））研究成果報告書）

角南北斗（2009）「Webデザインの実践プロセス 日本語教師ができること」東京外国語大学留学生日本語教育センター作文コーパス研究会（2）「教育分野におけるWebデザイン実践のプロセスについて」資料（2009年2月23日）

滝沢直宏（2007）「日本語研究とコーパス：日本語教育への応用を視野に」東京外国語大学留学生日本語教育センター FD 研修会におけるハンドアウト（2007 年 11 月 1 日）

『全学日本語プログラム 3 年間の報告書』（2007）東京外国語大学留学生日本語教育センター

『全学日本語プログラム履修案内』（2009 年度春学期、2009 年度秋学期）東京外国語大学留学生日本語教育センター

文部科学省「質の高い大学教育推進プログラム（教育 GP）」
（http://www.mext.go.jp/a_menu/koutou/kaikaku/gp/program/08033118.htm）