

LARP at SCU コーパスのウェブ検索エンジンの開発 と分析

陳相州/Chen Shiang-jou

東吳大學日本語文學系 副教授

Department of Japanese Language and Culture, Soochow University

【摘要】

本研究目的為讓更多人可以使用 LARP at SCU 語料庫，針對語料庫內學習者第一次所撰寫的作文資料，建置了網頁檢索引擎。此網頁檢索引擎除了能夠下載過去以來以 DVD 形式發行的 LARP at SCU 語料庫的內容之外，也讓使用者可以透過「檢索條件」、「關鍵字檢索」的功能進行更多樣化地運用。此外，本研究也利用此檢索引擎調查了台灣日語學習者在三年半的調查期間內使用在作文上的日語語彙，以釐清台灣日語學習者在日語語彙習得上的狀況。

【關鍵詞】

LARP at SCU 語料庫 檢索系統 台灣日語學習者 習得 語彙

【Abstract】

This study aimed to enable more people to access the “LARP at SCU” corpus by creating a web-based search engine for learners’ very first compositions in the corpus. This web-based search engine enables users to download old data from the “LARP at SCU” corpus that had been released in the form of DVDs in the past and provides functions such as “search criteria” and “keyword search” for more diversified applications. Moreover, this study also summarized the Japanese vocabularies applied in the compositions written by Japanese learners in Taiwan during the 3-and-a-half years of the survey, in order to identify the difficulties Japanese learners in Taiwan may face while acquiring this language.

【Keywords】

LARP at SCU corpus, online search system, Japanese learners in Taiwan,
acquiring, vocabulary

1. はじめに

LARP at SCU (Language Acquisition Research Project at Soochow University) は台湾人日本語学習者の日本語習得のプロセスを記録するために行ったプロジェクトである。このプロジェクトでは当時、東呉大学日本語学科に在籍した学習者 37 名が一年生の下学期から卒業までの間（2004 年 3 月 17 日－2007 年 5 月 16 日）、毎月書いた作文と録音データが収録されている。このプロジェクトの概要は次のようである¹。

このプロジェクトに参加した学習者は毎月決まった日時と場所で与えられたテーマについて辞書などを見ずに約 600 字の作文を書く。その後、書いた作文を朗読し、フォローアップ・インタビューを受ける。インタビュー完了後、学習者は同じテーマについてもう一度作文を書き、朗読する。学習者の作文、朗読とフォローアップ・インタビューの内容はすべて LARP at SCU コーパスに収録された。

LARP at SCU コーパスは台湾人日本語学習者の 3 年間半の日本語習得過程に産出した書き言葉と話し言葉を記録したもので、データの性質から言うとかかなり貴重な縦断的学習者データである。迫田（2002）が説明した通り、縦断的研究は少人数の特定の対象者に長期間にわたって行うものであるが、少人数のデータのため、結果を一般化することは難しい。だが、被験者のある一定期間の言語の実態を提示できる長所もあり、このような研究を積み重ねることにより、問題点の抽出、習得要因の焦点化が可能になるので、多面的な角度からの研究が望まれると指摘している。横断的研究に比べ、縦断的研究はかなり時間と労力をかけるものなので、あまり多くの研究がなされていない。一方、LARP at SCU コーパスには既に台湾人日本語学習者の 3 年間半の産出物が格納されており、データ収集、文字化などの手間がかかる作業を省くことができ、台湾人日本語学習者の習得状況を考察するにはかなり便利だといえよう。

従来、LARP at SCU コーパスは DVD の形で公開されているが、送料や郵送時間を考慮すると、海外にいる人にとって到底便利とは言えない。その上、DVD に入っているのは文字データのみで、検索ツールなどは入ってい

¹ LARP at SCU のプロジェクトについて、詳細は陳・許・王（2007）を参照。

ない。言い換えると、DVD を手に入れたらすぐ使える状態ではなく、閲覧や検索するのにやはり不便である。そこで、本研究では LARP at SCU コーパスを広く一般に利用してもらえるように、ウェブで一般公開し、多様な検索機能を備えた検索エンジンを開発した。また、その使用例の一つとして、台湾人日本語学習者が 3 年間半の調査期間で執筆した作文に出現した日本語の語彙を分析し考察した。

2. 学習者の日本語作文コーパス

この節では、LARP at SCU コーパスを含め、今既存している学習者の日本語作文コーパスについて説明し、本研究で構築したウェブ検索エンジンの意義を述べる。

2.1 公開中の学習者の日本語作文コーパス

今現在、公開されている主な学習者の日本語作文コーパスを以下で順次で紹介していく。

a. 「日本語学習者作文コーパス」²

このコーパスは早稲田大学の李在鎬先生と東京外国語大学の伊集院郁子先生³が収集した学習者の作文をコーパス化したものであり、中国語と韓国語を母語とする学習者、計 304 名の作文が収録されている。データの性質からいうと、横断的作文データ (cross-sectional composition data) である。ウェブ検索エンジンがあり、学習者の日本語レベルや正用・誤用からデータを検索することができる。また、このコーパスを利用した研究論文としては、李・宮岡・林 (2013) などが挙げられる。李・宮岡・林 (2013) は言語テストの得点と学習者が書いた作文のテキスト情報量の関連性について考察したものである。

² <http://sakubun.jp.org/>

³ 「日本・韓国・台湾の大学による日本語意見文データベース」であり、作文のデータもウェブで公開されている。

(<http://www.tufs.ac.jp/ts/personal/ijuin/terms.html>)

図1 「日本語学習者作文コーパス」のインターフェース

b. 「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」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」(以下、対訳作文 DB) に収録されたのは、23 国、計 1754 名の日本語学習者が書いた作文と彼らが母語で同じ内容を書いた作文である。横断的作文データであるが、すべての作文に母語話者の訂正があるわけではない。また、学習者の日本語レベルも明示されていなく、検索エンジンもなく、利用者にとっては便利とはいえない。曹 (2000) では、この対訳作文 DB に収録された学習者が日本語で書いた作文と母語である中国語で書いた作文を利用し、日中モダリティ表現を考察した。

⁴ http://contr-db.ninjal.ac.jp/essay_01.html

データファイルリスト		作文・添削データの簡単な説明					
	執筆者 ID	日本語作文 (txt)	手書き日本語作文 (jpg)	対訳作文 (txt)	対訳作文 (pdf)	手書き添削 (jpg)	添削 (xml)
一括選択		☐	☐	☐	☐	☐	☐
1	CN011	☐ JaZhCN011_01_1.txt 表示	☐ 2ページ リスト	☐ ZhZhCN011_01_1.txt 表示	☐ ZhZhCN011_01_1.pdf 表示		
2	CN012	☐ JaZhCN012_02_1.txt 表示	☐ 2ページ リスト	☐ ZhZhCN012_02_1.txt 表示	☐ ZhZhCN012_02_1.pdf 表示		

図2 「対訳作文 DB」のインタフェース

c. 「日本語教育のためのタスク別書き言葉コーパス」

このコーパスには日本語母語話者 (30 名)、韓国人日本語学習者 (30 名)、中国人日本語学習者 (30 名) が 12 のタスクについて書いた、計 1080 編の作文が収録され、書籍と一緒に販売されている。横断的的作文データであり、検索システムは備えていない。12 つのタスクの達成度により、学習者を上位群、中位群、下位群の 3 つのグループに分けている。このコーパスを利用した研究には趙 (2015) などがある。趙 (2015) は挨拶語、人称代名詞、指示代名詞などの観点から、中韓の留学生と日本人学習者との違いの一端を明らかにした。

d. 「台湾人日本語学習者コーパス」⁵

このコーパスには 1180 名の台湾人日本語学習者が書いた、計 1563 編の横断的的作文のデータが収録されている。誤用タグが付与されており、ウェブインターフェイスによる検索環境も用意されている。更に、学習者の属性とレベルなどの情報も付いている。

⁵ <http://corpora.flld.ncku.edu.tw/>

台灣日語學習者語料庫
台灣日語學習者コーパス
The Corpus of Taiwanese Learner of Japanese (CTLJ)
 以CTLJ為本之日語線上數位學習教材(Japanese e-Learning Based on CTLJ)
 線上日語寫作輔助學習系統(Online Supporting System for Japanese Composition)
 CTLJ誤用辭典(Japanese Misuse Based on CTLJ)

SEARCH

●日本語原文 ○日本語修正文 ○中国語 ○品詞

学習時間数: [不設限] 学科: [不設限] (学習時間数は「学習時間表」を参照してください。)

品詞限定: -2: [不設限] -1: [不設限] 0: [不設限] +1: [不設限] +2: [不設限] (機能を使用)

テーマ:

全て選択 全て取消 逆選択

2003學年度data

☒ うわさ話	☒ おすすめの料理と作り方
☒ ニュースを見て考えたこと	☒ 勉強について
☒ 夏休みの生活	☒ 思い出の旅行
☒ 私のお気に入りの店	☒ 私の夢
☒ 私の好きな映画	☒ 私の宝物
☒ 私の実家	☒ 自己紹介
☒ 高校時代の私	

2006-7學年度data

☒ ○○は○○すべきだ	☒ ナラティブ
☒ 思い出	☒ 社会問題

2008學年度data

☒ レポート	☒ 冬休みの生活
☒ 台湾と日本の違うところ	☒ 地球温暖化について考えたこと
☒ 家族との思い出	

作者資料:

学年: [不設限]

出身学校の種類? [不設限]

語学スクールや学習塾での学習経験の有無?
☒ 不設限
☐ 否
☐ 是

日本での居住経験の有無?
☒ 不設限
☐ 否
☐ 是

母語 (複数選択可)?
☐ 中文(國語)
☐ 閩南語
☐ 原住民語

日本語能力検定級数?
 級数: [不設限]

図3 「台湾人日本語学習者コーパス」のインタフェース

e. 「多言語母語の日本語学習者横断コーパス」⁶

「I-JAS (International Corpus of Japanese as a Second Language)」とも呼ばれ、異なった12言語を母語とする日本語学習者、計1000人が12のタスクについて書いた作文が収録されている。横断的的作文データである。学習者を実施した日本語能力テストの結果によって学習者のレベル判定があるため、ウェブでレベル別、母語別、学習環境別にデータの比較が可能である。但し、誤用情報は付与されていない。このコーパスを利用した研究としては、砂川(2018)などが挙げらる。砂川(2018)では学習者と母語話者の「てしまう」使用状況を調査し、中級以降で指導が必要な「てしまう」の用法を提言した。

⁶ <http://lsaj.ninjal.ac.jp/?cat=3>



図4 「多言語母語の日本語学習者横断コーパス」のインタフェース

f. 「LARP at SCU コーパス」

繰り返しになるが、LARP at SCU は Language Acquisition Research Project at Soochow University の略で、東呉大学日本語学科の学習者 37 名を対象とし、彼らが 1 年生後期から 4 年生後期まで、月に一回執筆した作文とインタビューを記録したプロジェクトである。また、このプロジェクトでは、学習者の日本語能力を知るために、学習者全員に対して SPOT (Simple Performance-Oriented Test) テストも実施した。LARP at SCU の記録物を DVD という形で公開されていた。

LARP at SCU コーパスを利用した研究が多く見られる。その研究は、日本語の条件副詞、漢語語彙、指示詞、副詞、条件文、外来語、焦点助詞「も・は・だけ・しか・さえ」、接続助詞「から」「ので」、中日対応語彙、否定表現「ない」、可能表現、外来語の特殊拍、「にとって」の使い方、自他動詞、授受表現、使役文といった様々なテーマにわたっている (盧 2007、陳 2007、後藤 2007、遠藤 2008、菊川 2010、羅 2010、瞿 2010、松倉 2010、陳 2011、馮 2011、金 2011、重光 2012、太田垣 2012、張 2013、李 2013、瞿 2014、呉 2014、黃 2015、楊 2016、胡 2016)。

2.2 本研究で構築したウェブ検索エンジンの意義

以上の 6 つの公開されている主な学習者の日本語作文コーパスを以下の表

1 にまとめる。

表 1 主な学習者の日本語作文コーパスの比較

学習者の日本語 作文コーパス	データ性質	誤用標記	学習者の日本語 レベル	ウェブ検索エン ジン
a. 「日本語学習 者作文コーパ ス」	横断	○	○	○
b. 「対訳作文 DB」	横断	△ ⁷	×	×
c. 「日本語教育 のためのタスク 別書き言葉コー パス」	横断	×	○	×
d. 「台湾人日本 語学習者コーパ ス」	横断	○	○	○
e. 「多言語母語 の日本語学習者 横断コーパス」	横断	×	○	○
f. 「LARP at SCU コーパス」	縦断	×	○	×

利用者の分析上の利便性を考え、誤用タグ付きで、学習者レベルも付与されており、オンラインで検索できる環境が一番いいと考えられる。しかし、表 1 からわかるように、今公開されている主な学習者の日本語作文コーパスの中には、誤用タグと学習者レベルの情報が付与され、またウェブ検索エンジンも用意されているのは、a. 「日本語学習者作文コーパス」と、d. 「台湾人日本語学習者コーパス」しかない。また、a. 「日本語学習者作文コーパス」と、d. 「台湾人日本語学習者コーパス」に収録されているのは学習者の横断

⁷ 一部のみ。

的作文データである。縦断的作文データが公開されているのは、本研究の検索エンジンの材料である f. 「LARP at SCU コーパス」 のみのようである。したがって、台湾人日本語学習者の縦断的作文データが収録されている LARP at SCU コーパスを対象とし、誤用・学習者レベル情報付きで、オンラインで検索できる本研究のウェブ検索エンジンの作成には意義があると考えられる。

前述したように、LARP at SCU コーパスは実に多くの研究に使用されている。そのため、本研究が構築した LARP at SCU コーパスのウェブ検索エンジンの公開を通し、もっと多くの研究に利用され、多分野にわたる研究が期待されよう。

3. 研究の目的

本研究では以下の 3 点を研究目的とする。

- (1) LARP at SCU コーパス内の作文データに誤用情報を付与する。
- (2) 多様な検索機能を備えた検索エンジンを開発し、ウェブで一般公開する。
- (3) 本研究で構築したウェブ検索エンジンを利用し、台湾人日本語学習者が 3 年間半の調査期間で書いた作文に出現した日本語の語彙の変化を調査する。

4. ウェブ検索エンジンの開発

4.1 前置作業

上述したように、LARP at SCU コーパスには学習者の作文、朗読とフォローアップ・インタビューが格納されているが、本研究では学習者が辞書を見ずに書いた最初の作文のデータを中心に、検索エンジンを作成してみた。その理由は、学習者が最初に書いた作文の方が学習者の日本語能力を忠実に反映しているものと考えられるためである。

調査期間で学習者が書いた作文のテーマと篇数は以下の通りである。対象となる作文数は計 953 篇である。

表 2 調査期間で書いた作文のテーマと篇数

回数	調査期間	学年	作文のテーマ	篇数
1	2004. 03. 17	1 年下学期	私の一日	35

2	2004. 04. 14		春休み	37
3	2004. 05. 05		私の部屋	36
4	2004. 06. 02		私の夢	36
5	2004. 06. 23		高校生活	31
6	2004. 09. 22	2 年上学期	忘れられない出来事	33
7	2004. 10. 27		十年後の私	34
8	2004. 11. 24		もし一千万円当たったら	35
9	2004. 12. 08		大学生活に期待すること	32
10	2005. 01. 05		私と日本語の出会い	33
11	2005. 03. 02	2 年下学期	お正月	33
12	2005. 04. 06		携帯電話	29
13	2005. 05. 04		母の日	29
14	2005. 05. 25		友情	29
15	2005. 06. 08		流行	29
16	2005. 09. 28	3 年上学期	夏休み	27
17	2005. 10. 19		私の愛用品	28
18	2005. 11. 23		旅する	27
19	2005. 12. 24		選挙	26
20	2006. 01. 04		2006 年を迎えて	26
21	2006. 03. 01	3 年下学期	最近の出来事から	26
22	2006. 03. 29		スポーツ	26
23	2006. 04. 26		町	25
24	2006. 05. 24		私の愛読書	24
25	2006. 06. 14		最後の夏休み	24
26	2006. 09. 27	4 年上学期	台湾のデモについて	26
27	2006. 10. 18		ゴミ問題	27
28	2006. 11. 22		台湾の外食文化	26
29	2006. 12. 13		コーヒー文化	26
30	2007. 01. 03		台湾の野良犬	26
31	2007. 03. 07	4 年下学期	少子化	25
32	2007. 04. 11		大学生の恋愛観	24

LARP at SCU コーパスのウェブ検索エンジンの開発

33	2007. 05. 16		LARP に参加した感想	23
合計				953

本研究がウェブ検索エンジンを構築する前に、まず、データの正確さを確保するため、LARP at SCU コーパスに入っているすべてのデータをもう一度学習者の手書きした作文の PDF ファイルと照らし合わせ、確認・修正作業を行った。次に、本研究では LARP at SCU コーパスに格納されている学習者が最初にした計 953 篇の作文を 1 名の日本語母語話者教師⁸に添削してもらった。添削は教師の手書きによるものであるため、教師による添削情報の電子化作業も行った。

4.2 ログイン・インターフェイス

LARP at SCU コーパスを学習者から研究者まで幅広い利用者が容易に利用できるように、本研究が構築したウェブ検索エンジンを「<https://sousyuuus.herokuapp.com/Grammar>」のドメインで一般公開している。以下はウェブ検索エンジンに入る前のインターフェイスである。


LARP at SCU線上搜尋系統

・ 説明：

此系統可以檢索LARP at SCU語料庫內所收錄的資料以及日語母語話者的修正內容。關於LARP at SCU語料庫的詳細說明請參照以下網址。

🔗 http://web-ch.scu.edu.tw/japanese/web_page/3936

・ 謝辭：

此系統為科技部計畫案「建置台灣日語學習者縱斷性寫作之誤用資料庫」（MOST106-2410-H-031-043-）與「LARP at SCU語料庫支線上檢索平台建置及其應用研究」（MOST107-2410-H-031-047-）的研究成果之一。

図 5 ウェブ検索エンジンのインターフェイス

利用者が LARP at SCU のプロジェクトについてもっと詳しく知りたいで

⁸ 人によって添削基準が異なる可能性があるため、本研究では 1 名の日本語母語話者に添削の依頼をした。

あれば、画面上のリンクをクリックすることで、以下のような東呉大学日本語学科の公式サイトにある LARP at SCU コーパスの説明ページに繋がることができる。

關於LARP at SCU語料庫

代表:陳 淑娟

歡迎各位來到LARP at SCU語料庫網站，首先容我介紹一下本語料庫的建立經過，同時竭誠歡迎對日語習得研究有興趣的同好，充分利用本語料庫，進入語言習得研究的世界。

LARP at SCU是Language Acquisition Research Project at Soochow University的略稱，而Soochow University是(台灣)東吳大學的英文名稱，是東吳大學日本語文學系老師與研究生們爲了研究「學生的日語成長」而啟動的研究團體，由我發起，結合本系專兼任同好的老師與研究生數十名，於2004年3月開始策劃，執行「日語習得研究」，稱為LARP at SCU。

図6 LARP at SCU コーパスの説明ページの一部

黄(2009)が構築した「台湾人日本語学習者コーパス」はシステムが破壊されるのを防ぐために、システムに入るには認証を必須とされている。また、Session を用いて利用者登録の際の個人資料と合わせ、利用者の身分を判断して操作権限を行っているという。しかし、外部からのサイバー攻撃によって利用者が登録した個人情報漏えいしてしまうという事件をよく耳にしたため、本研究ではこの懸念を解消するために、あえて利用者登録をせずに誰でも利用可能な環境を作ることにした。

本研究では利用者の使用上の便宜を考慮し、ウェブ検索エンジンの利用のほかに、従来DVDの形で発行したLARP at SCU コーパスの内容をそのまま一括してダウンロードすることも可能にした。但し、本研究が開発したウェブ検索エンジンを利用する、またはLARP at SCU コーパスを一括してダウンロードするには、以下のような利用規約に同意せねばならない。

■使用條款

在使用LARP at SCU線上搜尋系統之前，您必須瞭解並同意遵守以下使用條款之內容。

1. 僅能使用在教學與學術研究上。

2. 禁止複製LARP at SCU線上搜尋系統之內容。

3. LARP at SCU線上搜尋系統內之資料皆有著作財產權。

4. 謹慎處理個人隱私之相關資料。

5. 以論文等形式公開發表時，請務必註明使用LARP at SCU線上搜尋系統之聲明。

若未遵守以上要點而產生之問題，該使用者須承擔全部責任。

図 7 本研究が構築したウェブ検索エンジンの利用規約

4.3 検索インターフェイス

本研究が開発した LARP at SCU コーパスのウェブ検索エンジンには検索条件とキーワード検索の機能が付いている。検索インターフェイスの画面は以下の通りで、検索条件とキーワード検索の機能については次の小節で詳しく説明する。

LARP at SCU線上搜尋系統

原文關鍵字搜尋

關鍵字

☒ 原文 ☐ 訂正文 ☐ 誤用品詞

搜尋選項1

學習者 ☐ All

☐ L1

☐ L2

☐ L3

☐ L4

☐ L5

☐ L6

☐ L7

☐ L8

☐ L9

☐ L10

☐ L11

☐ L12

☐ L13

☐ L14

☐ L15

☐ L16

☐ L17

☐ L18

☐ L19

☐ L20

☐ L21

☐ L22

☐ L23

☐ L24

☐ L25

☐ L26

☐ L27

☐ L28

☐ L29

☐ L30

☐ L31

☐ L32

☐ L33

☐ L34

☐ L35

☐ L36

☐ L37

日文等級 ☐ All (日文等級說明)

☐ 下位群

☐ 中位群

☐ 上位群

搜尋選項2

年級 ☐ All

☐ 一年級下	☐ 二年級上	☐ 二年級下	☐ 三年級上
☐ 1 私の一日	☐ 6 忘れられない出来事	☐ 11 お正月	☐ 16 夏休み
☐ 2 春休み	☐ 7 十年後の私	☐ 12 携帯電話	☐ 17 私の愛用品
☐ 3 私の部屋	☐ 8 もし一千万円当たったら	☐ 13 母の日	☐ 18 旅する
☐ 4 私の夢	☐ 9 大学生生活に期待すること	☐ 14 友情	☐ 19 選挙
☐ 5 高校生活	☐ 10 私と日本語の出会い	☐ 15 流行	☐ 20 2006年を迎えて
☐ 三年級下	☐ 四年級上	☐ 四年級下	
☐ 21 最近の出来事から	☐ 26 台湾のデモについて	☐ 31 少子化	
☐ 22 スポーツ	☐ 27 ゴミ問題	☐ 32 大学生の恋愛観	
☐ 23 町	☐ 28 台湾の外食文化	☐ 33 LARPに参加した感想	
☐ 24 私の愛読書	☐ 29 コーヒー文化		
☐ 25 最後の夏休み	☐ 30 台湾の野良犬		

搜尋

重新搜尋

図8 ウェブ検索エンジンの検索インターフェイス⁹

4.3.1 検索条件

本研究が開発したウェブ検索エンジンの検索条件には「検索条件 1」と「検索条件 2」の 2 つがある。「検索条件 1」は執筆した学習者自身に対する検索条件であり、「検索条件 2」は作文テーマに対する検索条件である。

「検索条件1」には「学習者」と「日本語レベル」という検索条件がある。

「学習者」という検索条件では、利用者が学習者を指定し、その学習者の調査期間で書いた作文を閲覧することが可能で、特定の学習者に対して縦断的研究を行う際に有益である。または、「ALL」というボタンをクリックし、学習者を全員選択し、ある時期の学習者たちの日本語の使用状況を観察することもできる。

「日本語レベル」という検索条件では、異なった日本語能力を持つ学習者同士の習得状況を比較することが可能となった。「日本語レベル」という検索条件は当初、LARP at SCU プロジェクトに参加した 37 名の学習者を対象に行った SPOT テスト¹⁰（一年生下半期）の結果によるものである。本研究

⁹ 学習者のプライバシーを保護するために、本研究が構築したウェブ検索エンジンには「L+番号」の形式で個々の学習者を示す。学習者のLはlearnerを表し、番号は通し番号である。

¹⁰ 小林（1996）が開発した「日本語学習者に対するプレースメントテストとしての SPOT（Simple Performance-Oriented Test）」のカセットテープを使用した。

ではその成績に基づき、学習者を上位群、中位群、下位群という三つのグループに分けてみた。SPOT テストは 60 点が満点となるテストである。このテストの学習者全員の平均は 42 点で、標準偏差は 10.01 点であった。人数の分布を考慮し、平均点を基準としてその前後 5 点、すなわち 38 点から 46 点の学習者のデータを中位群とし、47 点以上の学習者を上位群、37 点以下の学習者を下位群にした。その結果、上位群に分類された学習者は 13 名（平均値：52.00、標準偏差：3.00）、中位群に分類された学習者は 12 名（平均値：42.42、標準偏差：2.19）、下位群に分類された学習者は 11 名（平均値：29.73、標準偏差：6.20）となった。以下ではこの 3 つのグループの成績分布と人数を示す。

表 3 3 つのグループの成績分布と人数

	成績分布	人数
上位群	47-57	13
中位群	38-46	12
下位群	21-37	11

一方、「検索条件 2」は調査期間内のすべてのテーマを選択することはもちろん、テーマ別もしくは学期別の選択も可能となった。

4.3.2 キーワード検索

本研究が開発したウェブ検索エンジンのキーワード検索の機能には「原文」、「訂正文」と「誤用品詞」がある。利用者は学習者が書いた日本語の「原文」、または教師の添削が行われた「訂正文」から、キーワードを検索することができる。「誤用品詞」は日本語母語話者教師が添削した箇所の品詞を意味している。

本研究では「誤用品詞」を判断するために「Web 茶まめ」¹¹を利用した。「Web 茶まめ」は日本語形態素解析器 Mecab によって形態素解析を行うものである。本研究では形態素解析辞典を「IPAdic（現代語）」に指定して形態素解析を行った。本研究の「誤用品詞」には「連体詞」、「接頭詞」、「名詞」、

¹¹ <http://chamame.ninjal.ac.jp>

「動詞」、「形容詞」、「副詞」、「接続詞」、「助詞」、「助動詞」、「感動詞」、「記号」、「フィラー」、「未知語」のほかに、「×」という記号もある。「×」という記号が付けられたのは、日本語母語話者教師が学習者の書いたものをすべて削除した箇所である。よって、本研究が開発したウェブ検索エンジンの「誤用品詞」には「×」も含め、計 14 の種類がある。利用者が下の図に示したボタンをクリックすると、「誤用品詞」のセレクトボックスが出てくる。このようにして利用者が考察したい品詞だけを選んで検索することができる。



図 9 誤用品詞の一覧を開くボタン

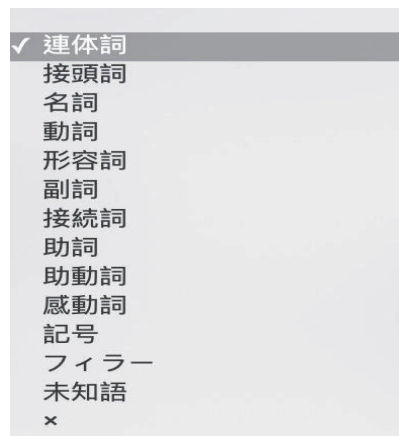


図 10 品詞を選択するセレクトボックス

なお、誤った操作を防ぐために、「検索条件 1」と「検索条件 2」を選択しない状態では検索アイコンが利用できないように設定した。

4.3.3 検索結果の表示

利用者の指定した検索条件と入力した検索内容に応じ、システムからその検索結果が返される。その検索結果以外に、「学習者番号」、「日本語レベル」、「学年情報」、「何回目（の作文）」、「原文」、「訂正文」と「誤用品詞」も含

まれている。以下では学習者 L1 が書いた「私の部屋」の作文内容から誤用品詞を「名詞」に指定した検索結果である。

4筆記録, 目前是 0~4

学習者	等級	年級	回数	原文	訂正文	誤用品詞
L1	上位群	一年生下	3	部屋の大きさはだいたい六畳で、中にはベットや机や本棚やステレオなどがあります。	部屋の大きさはだいたい六畳で、部屋にはベッドや机や本棚やステレオなどがあります。	副詞 名詞 名詞 名詞
L1	上位群	一年生下	3	窓から見える景色は庭の庭と青い空です。	窓から見える景色は、隣の庭と青い空です。	記号 名詞
L1	上位群	一年生下	3	部屋の両側は六人つづのベットセットが置いてあります。	部屋の両側には六人のベッドセットが置いてあります。	助詞 名詞
L1	上位群	一年生下	3	上はベット、下はたんすと本棚と机があります。	上はベッド、下はたんすと本棚と机があります。	名詞

図 11 検索結果の表示例

図 11 からわかるように、「誤用品詞」の部分に利用者が指定した「名詞」のみが黄色く示されている。また、同時に「名詞」に対応する箇所を「原文」で赤い色で表れ、「訂正文」で青い色で表示されている。なお、利用者が検索結果を自由に運用してもらえるように、本研究が開発したウェブ検索エンジンでは検索結果をコピーし、表計算ソフトに貼り付けることも可能にした。

5. 分析と考察

この節では、本研究で構築した LARP at SCU コーパスのウェブ検索エンジンを利用し、その 1 つの例として台湾人日本語学習者が 3 年間半の日本語学習過程で書いた作文に出現された日本語の語彙を分析してみる。また、学習者が大学に入った時点では、日本語の既習者と未習者がおり、同じ大学一年生とは言え、日本語レベルは全く同じとは限らない。そのため、本研究では学習初期で異なった日本語能力をもつ学習者がこの 3 年間半の日本語学習過程で作文に産出した日本語の語彙の変化も注目する。

Nation (2007) が説明した通りに、学習者の語彙知識の全体像を掴むために、学習者によって語彙がどのように使われるのかを調べる必要がある。そこで、本研究では台湾人日本語学習者が 3 年間半の日本語学習過程で毎回の

作文に使用した日本語の語彙の総語彙数と語彙の種類を考察する。

LARP at SCU コーパスには 37 名の学習者のデータが格納されているが、実際では 3 年間半のデータがすべて揃っているのは 17 名だけであった。前節でも説明したように、LARP at SCU のプロジェクトが始動された学習者の 1 年生下学期の最初のときに、プロジェクトに参加した学習者に対して満点 60 点の SPOT テストが行われた。本研究ではその成績に基づき、上位群、中位群、下位群の三つのグループを分けた。この 17 名の学習者のグループ分けの結果を下の表で示す。

表 4 17 名の学習者のグループ分けの結果

グループ	学習者	人数
上位群	L20、L24、L27、L30、L34	5 名
中位群	L5、L8、L9、L10、L14、L21、L25、L26、L29	9 名
下位群	L6、L28、L35	3 名

以下の分析はこの 17 名の学習者を対象とし、彼らが 3 年間半の日本語学習過程で執筆した作文の中に出現した総語彙数と語彙の種類を分析し考察する。

5.1 総語彙数

台湾人日本語学習者が 3 年間半の調査期間で書いた作文の総語彙数の変化を調査するにあたって、上位群、中位群、下位群の学習者が毎回の調査で書いた作文の語彙を抽出し、グループごとに毎回の調査で書いた作文の延べ語数の平均値を計算した。その結果は下の図 12 の通りである。

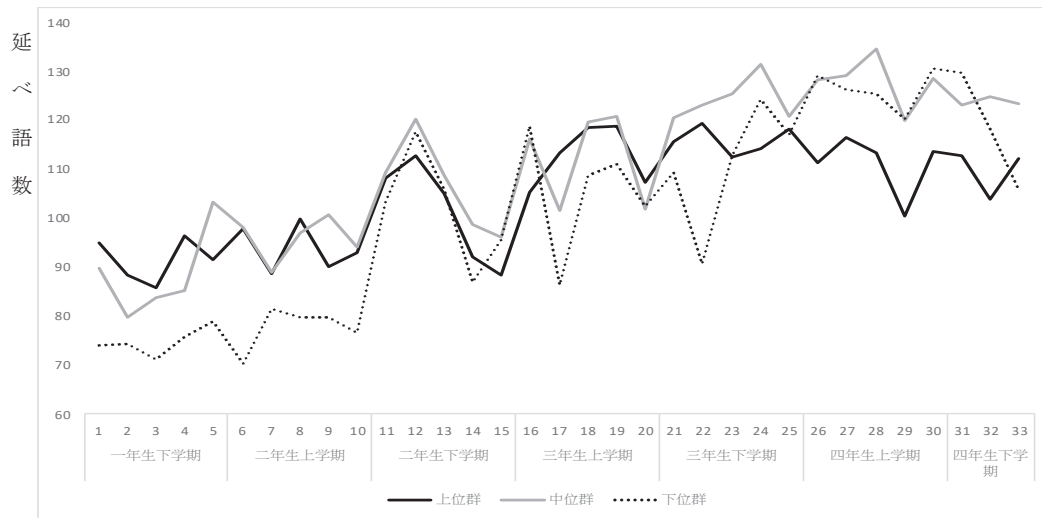


図 12 三つのグループの延べ語数の推移

各グループの延べ語数の推移状況を詳しく把握するために、各グループの延べ語数の推移の近似曲線を出してみた。図 13 に示した回帰式「 $y=ax+b$ 」の a は曲線の傾きであり、直線の傾斜の具合を示すものである。回帰式「 $y=ax+b$ 」の a の数値が大きいほど、直線の傾斜が激しく、調査期間での成長が著しいと考えられる。

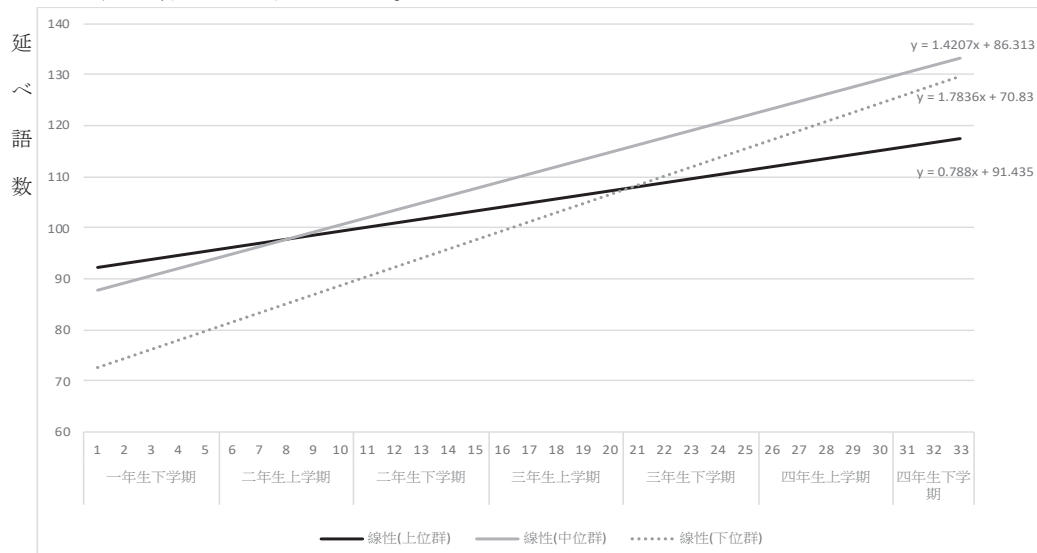


図 13 三つのグループの延べ語数の近似曲線の傾き

図 13 からわかるように、上位群の学習者は一年生のときに執筆した作文の語彙数が一番多かったが、二年生に入り、まず中位群の学習者に超えられ、更に三年生下学期のときに下位群の学習者に超えられた傾向が見られた。また、この 3 年間半の調査期間で作文に現れた総語彙数の成長がもっとも著しいのは下位群の学習者であった。

5.2 語彙の種類

各グループの学習者の作文に出現した語彙の種類を究明するために、本研究では各グループの学習者が調査期間で書いた作文を解析し、毎回の作文に現れた異なり語数を集計した。その結果は図 14 で示す。

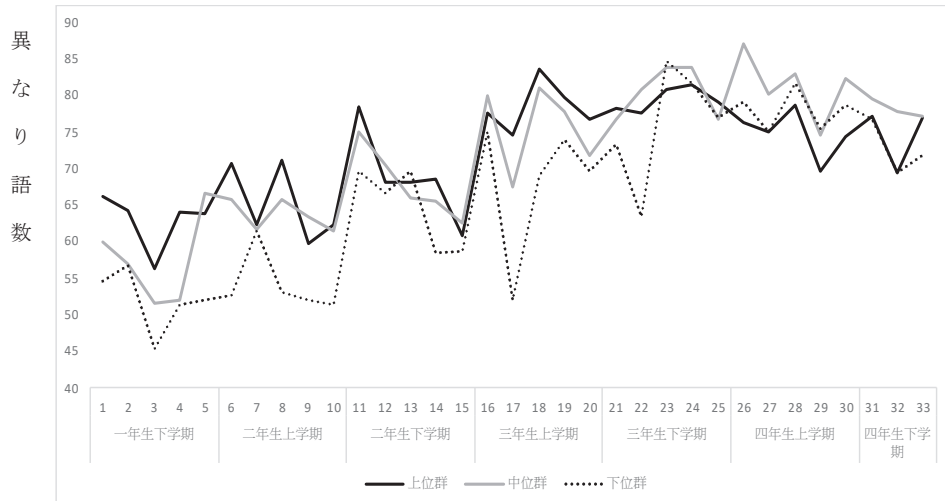


図 14 三つのグループの異なり語数の推移

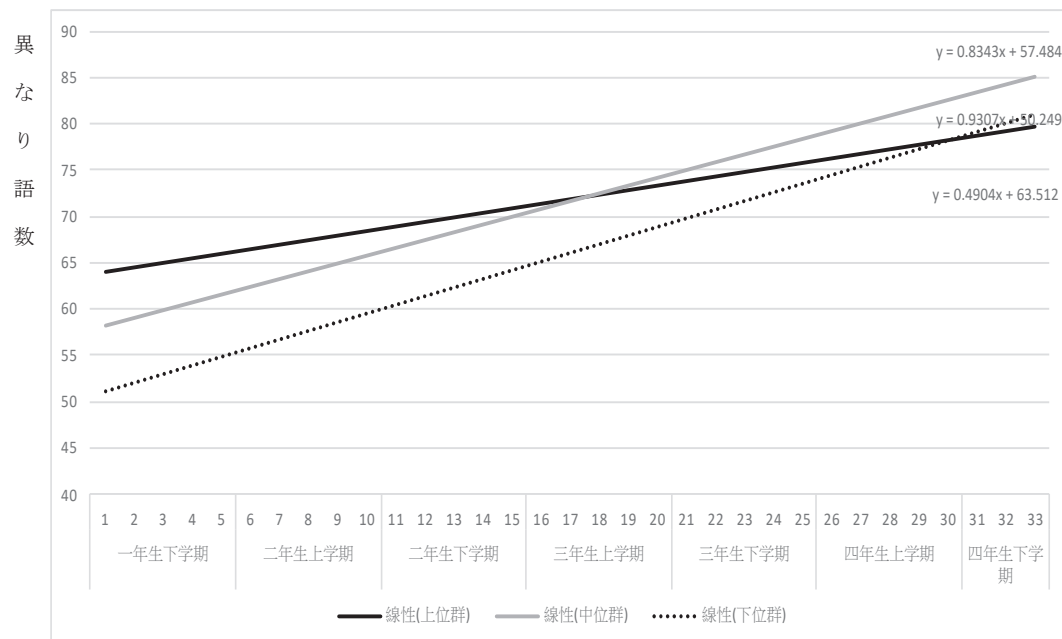


図 15 三つのグループの異なり語数の近似曲線の傾き

上位群学習者、中位群学習者と下位群学習者が調査期間で書いた作文の異なり語数の推移を詳しく観察するために、前節と同様に、図 15 のような近似曲線を出してみた。図 15 に示したように、上位群の学習者は一、二年生のときに作文に使用した語彙の種類がもっとも多かった。しかし、三年生と四年生のときに、中位群学習者と下位群学習者に比べ、作文に用いた語彙の種類はかえって少なかったような傾向が窺われる。また、近似曲線の傾きを見ると、この 3 年間半の調査期間では下位群学習者の成長がもっとも顕著であるということもわかった。

5.3 考察

以下では、3 年間半の調査期間で台湾人日本語学習者の作文に出現した語彙の総語数と種類の分析結果をまとめる。

LARP at SCU のプロジェクトでは学習者が 1 年生下学期の最初の回に SPOT テストを実施した。本研究ではその結果に基づき、プロジェクトに参加した台湾人日本語学習者を上位群、中位群、下位群の三つのグループに分

けた。上位群にグループ分けした学習者は全員、大学に入る前に日本語を学習した経験のある人である。本研究の調査を通してわかったことは次のようである。

作文に使用した日本語の語彙の総語数と種類においては、上位群学習者がほかのグループより多かったのは1、2年生の学習の初期段階のみであった。この3年間半の調査期間では特に、下位群学習者が作文に使用した語彙の総語数と種類に顕著な成長を見せ、卒業する直前に入学する前に日本語の学習経験のある上位群学習者に比べても遜色がないほどであった。本研究ではこの17名の台湾人日本語学習者を対象とし、3年間半の学習期間に作文に産出した日本語語彙の総語数と種類を追跡した結果、大学に入る前に日本語学習の経験があるか否かは作文の語彙の産出に影響があったのは、大学の初期段階のみであり、大学の後期段階に入ってその影響はもう見えないということが明らかになった。

6 おわりに

学習者が学習期間で産出した日本語はどのように変化するのかを明らかにするために、縦断的研究の成果を多く積み重ねる必要があると考えられる。LARP at SCU コーパスには37名の台湾人日本語学習者の3年間半の学習期間の産出物が格納されており、台湾人日本語学習者の習得状況を解明するには大変貴重なコーパスだといえよう。本研究ではLARP at SCU コーパスをより多くの人に使用してもらうために、ウェブ検索エンジンを開発し公開した。このウェブ検索エンジンは従来DVDの形で発行したコーパスの内容をそのまま一括してダウンロードすることができるほか、「検索条件」と「キーワード検索」の機能を通じてデータを多様に活用することもできるようにした。

本研究ではその使用例の一つとして、台湾人日本語学習者が3年間半の学習期間で作文に使用した日本語の語彙を調査した。語彙の総語数と種類を考察した結果、大学に入る前に日本語の学習経験のある学習者はほかのグループより語彙の総語数と種類が多く観察できたのは学習の初期段階のみであり、大学の後期段階に入っては既習者の優位性はもう観察できなくなってしまう。しかし、本研究の分析に使用した上位群・中位群・下位群の人数はかなり少ないため、出した結果を一般化することは難しいということは否めな

い。

LARP at SCU コーパスに収録されているのはかなり貴重な縦断的学習者データであるが、今後、この特徴を生かし、実際の学習者の作文を挙げて質的に比較検討していく。また、日本語を母語とする利用者の作業しやすい環境を作るために、本研究が構築したウェブ検索エンジンの日本語のインターフェイスを実装する作業も行う予定である。

参考文献

- 遠藤保代 (2008)『台湾人学習者における副詞の習得研究-作文データをもとに-』東呉大学日本語学科修士論文
- 太田垣良孝 (2012)『台湾人日語學習者在來語特殊拍書寫上之誤用分析-以 LARP at SCU 為例之通時性研究-』国立高雄第一科技大学応用日本語学科修士論文
- 菊川秀夫 (2010)「LARP コーパスに見る台湾人日本語学習者の産出語彙の変化」『銘傳日本語教育』13、pp. 206-229.
- 金裕文 (2011)『台湾人日本語学習者の可能表現に関する縦断的習得研究-LARP at SCU の作文データを中心に-』東呉大学日本語学科修士論文
- 黃淑妙 (2009)『日本語習得の達成度分析-「台湾人日本語学習者コーパス」(CTLJ)の構築と分析を中心に-』致良出版社
- 黃瑞穎 (2015)『探討台灣日語學習者對自他動詞習得的研究-LARP at SCU 作文語料庫的事例研究-』東呉大学日本語学科修士論文
- 胡君平 (2016)「台湾人学習者による日本語使役文の用法別の使用実態：LARP at SCU の分析結果から」『日本語教育』163、pp. 95-103.
- 後藤晃 (2007)『日本語指示詞についての事例研究-LARP at SCU の初級学習者における縦断的調査を中心に-』東呉大学日本語学科修士論文
- 小林典子、フォード丹羽順子、山本啓史 (1996)「日本語能力の新しい測定法<SPOT>」『世界の日本語教育』6、pp. 201-218.
- 吳旻樺 (2014)『台籍日語學習者之日語條件句型學習研究-以 LARP at SCU 為中心』東呉大学日本語学科修士論文
- 迫田久美子 (2002)『日本語教育に生かす第二言語習得研究』アルク
- 重光雨青 (2012)「台湾人学習者の作文に見られる『実は』の誤用分析-LARP at SCU のデータに基づいて」『台灣日語教育學報』19、pp. 141-168.

- 砂川有里子 (2018) 「中級以降で指導が必要なテシマウの用法について—学習者と母語話者の使用状況調査に基づく考察—」藤田保幸・山崎誠編『形式語研究の現在』 pp. 479-499.
- 曹大峰 (2000) 「作文コーパスによる日中モダリティ表現の対照研究—概言と確言—」前田洋著『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研究成果報告書』 pp.72-80.
- 翟翠翎 (2010) 「取り立て詞の縦断的習得研究」『東吳日語教育學報』35、pp. 50-69.
- 翟翠翎 (2014) 「作文から見る台湾人 JFL 学習者の取り立て詞の習得—縦断的研究—」『東吳日語教育學報』 43、pp. 44-73.
- 趙海城 (2015) 「上級—超級日本語学習者の作文から見た言語産出実態」『第7回コーパス日本語学ワークショップ予稿集』 pp. 293-302.
- 張秀玲 (2013) 「台湾人日本語学習者の『ニトッテ』の使用について : Larp at SCU コーパスを例に」『明海日本語』 18、pp. 55-64.
- 陳淑娟 (2007) 「作文における漢語語彙の習得についての考察—LARP at SCU のデータに基づく事例研究—」『台灣日本語文學報』 22、pp. 379-403.
- 陳淑娟・許夏珮・王淑琴 (2007) 「学習者言語のコーパスの構築に向けて—LARP at SCU を例として—」『銘傳日本語教育』 10、pp. 147-167.
- 陳淑娟 (2011) 「作文における『中日対応関係語彙』の習得についての考察—LARP at SCU コーパスの漢字語彙の使用から」『東吳外語學報』 33、pp. 25-45.
- 馮寶珠 (2011) 「作文における否定『ない』の習得についての考察—LARP at SCU の初級データに基づいて」『東吳外語學報』 32、pp. 51-76.
- 松倉朋子 (2010) 『台灣日語學習者之接續助詞「kara」和「node」之習得研究—以 LARP at SCU 語料庫為探討資料—』東吳大学日本語学科修士論文
- 楊淑惠 (2016) 『台灣人日本語學習者における授受表現の習得に関する一考察—LARP at SCU コーパスのデータを中心に—』東吳大学日本語学科修士論文
- 羅濟立 (2010) 「カタカナ語の習得についての事例研究—LARP at SCU による縦断的資料の分析から—」『台灣日本語文學報』 27、pp. 219-240.
- 李虹儀 (2013) 『台灣日語學習者之焦點助詞“dake”和“shika”的習得研究—Larp at SCU 作文資料為中心—』東吳大学日本語学科修士論文

- 李在鎬・宮岡弥生・林炫情(2013)「学習者コーパスと言語テスト:言語テストの得点と作文のテキスト情報量の関連性」『言語教育評価研究(AELE)』3、pp.22-31.
- 盧月珠(2007)「LARP at SCUにおける状態副詞の習得についての考察-初級から中級までの作文データを調査対象に」『台灣日本語文學報』22、pp. 431-455.
- Nation, I. S. P. (2007) Fundamental issues in modeling and assessing vocabulary knowledge. In H. Daller, J. Milton, & J. Treffers Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge*. pp. 93-115.

付記

本研究の成果は 106 年度科技部專題研究「建置台灣日語學習者縱段性寫作之誤用資料庫(MOST106-2410-H-031-043-)」、107 年度科技部專題研究「LARP at SCU 語料庫之線上檢索平台建置及其應用研究 (MOST107-2410-H-031-047-)」と東吳大學推動人工智慧應用研究「D-03 利用 LARP at SCU 語料庫的自然語言處理」によるものである。

本論文於 2020 年 3 月 11 日 到稿・2020 年 6 月 1 日通過審查。