Japanese Students' L1 Story Writing Corpus (JASWRIC): A New Dataset for Analysis of L1/L2 Japanese

「小中高大生による日本語絵描写ストーリーライティングコーパス」（JASWRIC）の構築：L1/L2 日本語研究の新しい資料として

石川慎一郎（神戸大学）
友永達也（神戸大学附属小学校），大西遼平，岡本利昭，勝部尚樹，川嶋久予，
岸本達也，村中礼子（神戸大学附属中等教育学校）

Japanese Students' L1 Story Writing Corpus (JASWRIC): A New Dataset for Analysis of L1/L2 Japanese

Shin'ichiro ISHIKAWA (Kobe University)
Tatsuya TOMONAGA (Kobe University, Elementary School), Ryohei ONISHI, Toshiaki OKAMOTO, Naoki KATSUBE, Hisayo KAWASHIMA, Tatsuya KISHIMOTO, Reiko MURANAKA (Kobe University, Secondary School)

要旨

本稿は、「小中高大生による日本語絵描写ストーリーライティングコーパス」（JASWRIC）の構築過程と概要を報告する。JASWRIC には、700 名の小中高大生による約 13.6 万語（短単位）の L1 日本語作文が収録されている。全データは、ダウンロード版とオンライン版（JASWRIC *Online*）の 2 系統で公開される。一般公開されている L1 の子どもの作文コーパスがほとんどない中で、JASWRIC は、L1 日本語の発達過程を調べる有益な資料となるだろう。また、JASWRIC のデータは、「多言語母語の日本語学習者横断コーパス」（I-JAS）で採用されたストーリーライティングのプロンプトを使って集められている。このため、JASWRIC は、I-JAS と併用することで、L1/L2 対照研究の参照データとしても使用可能である。

Abstract

This article introduces *Japanese Students' L1 Story Writing Corpus* (JASWRIC), which includes approximately 136,000-word L1 Japanese writing by 700 primary school, secondary school, and college students. The collected data is available as a download version as well as from the online corpus query system called JASWRIC *Online*. Considering that there have been almost no L1 Japanese students' essay corpora publicly available, we could expect that JASWRIC will be used as a valuable resource for the study of the development of children's L1 Japanese knowledge and skills. In addition, JASWRIC can also be utilized as reference data in L2 Japanese acquisition studies because its data is collected by the picture-based story writing task common to that adopted in *International Corpus of Japanese as a Second Language* (I-JAS). Adoption of the common picture prompts guarantees the validity in a contrastive study of L1 and L2 Japanese based on I-JAS and JASWRIC.

１．はじめに

「多言語母語の日本語学習者横断コーパス」（I-JAS）（迫田ほか、2020）がリリースされたことで、学習者による L2 日本語の習得過程については、大型コーパスを用いた実証的研究が広く行われるようになってきた。一方、母語話者の子ども（本研究では、広く未成年者を指す）の L1 日本語の発達過程を直接の研究対象としたコーパス研究は、管見の限り、必

ずしも多くない。これには、研究基盤となる公開コーパスの少なさが影響している。

もちろん、過去において、子どもによるL1日本語の産出データの収集例がなかったわけではない。まず、話し言葉について言えば、Talk Bank Systemのサブモジュールと位置付けられるCHILDES（Child Language Data Exchange System）に、幼児のL1日本語発話等を集めた12種の資料が掲載されている。ただ、PaidoJapanese資料（2〜5歳児85名の発音）とOkayama資料（2〜4歳児130名と親との対話）を除くと、大半は1名ないし2〜6名程度の小規模なデータである。これらは、音声学、発達心理学、会話分析など、特定の関心に即して集められた小規模データであり、一般的な意味でのコーパスとみなすことはむつかしい。ただ、近年になって、子どもの発話の体系的な収集も始まっており、小磯ほか（2020）では、8世帯10名の子どもの会話と、幼稚園に通う園児の会話を集める「子どもの会話コーパス」の開発計画が紹介されている。研究の成果が期待されるところである。

次に、書き言葉について言うと、データ収集の対象は、文字の習得が前提になるため、必然的に、小学生以降となる。小学校では作文を書かせる指導が広く行われており、データ自体は随所に存在しているわけだが、言語学ないし言語習得研究の目的でそれらを体系的に収集し、本人および保護者の許諾を得て公開の研究資料とすることのハードルは高い。とくに、小学校の国語授業で書かせることが多い自由作文には、子どものプライバシー情報が出やすいという問題もある。

初期の作文収集事例として、島村（1987）は、千葉県の3つの小学校に在籍する2年生・4年生・6年生約360名が「わたしの学校」「先生」「ともだち」というテーマで書いた作文を集めた。その一部（約34万語）は、後に「学校課題作文コーパス」として電子化されている（今田・宮城、2020）。国立国語研究所（1989）は各地の教育委員会などの編纂による小学生文集10種（10年分）から作文47万語を集めた。また、成田ほか（1995）は、小学校1年生から高校2年生と、大学生および大学留学生520名が「手」という同一のテーマで書いた「同題作文」を収集した。本資料を用いた研究としては、成田ほか（1995）に加え、村上・田中（1997）および田中（1998）がある。

近年では、より体系的な作文データの収集が試みられている。2010年代前半に作られた各種作文コーパスを紹介した冨士原ほか（2016）のまとめによれば、永田ほか（2010）は小学校5年生児童の読書ブログ約4万語を集め、坂本（2010）は全国の小学校のウェブサイトに公開されていた作文約1万本、約123万語を集め、鈴木ほか（2011）は中学校1年生から高校2年生の作文25万語を集めた。また、阿部ほか（2017）は、上述の成田ほか（1995）で集められたデータのうち、小中学生が書いた約5万語の作文を電子化し、あわせて、同じ題目に基づいて現代の小中学生が書いた作文約23万語を集め、総計28万語の「『手』作文コーパス」を構築した。最近の特筆すべき成果として、宮城・今田（2018）は、小学校1年生から中学3年生までの児童・生徒が「ゆめ」および「ぼくの／私のがんばったこと」というテーマについて授業内で書いた作文約5,300編を集め、約164万語の「児童・生徒作文コーパス」を構築した（作文数・語数は今田（2020）による）。なお、本コーパスは、冨士原ほか（2016）において70万語の「小・中学生通年作文コーパス」として言及されていたものである。本コーパスについては、作文中の「問題例」の抽出を可能にする「問題例検索システム（仮称）」が開発中とのことである（砂川、2020）。

こうした過去の作文収集はいずれも価値ある研究実践と言えるが、(1)小学生・中学生の作文が大半で、高校生以上の作文を含めた資料が少ない、(2)L2研究との接合性を考慮した資料が少ない、(3)緩やかなテーマで書かせた自由作文が多く、内容を統制して集めた資料

が少ない、(4)一般公開されているコーパスがほとんど存在しない、といった制約もある。この意味において、坂本が 2010 年時点で述べた「日本の子供の書き言葉コーパスは非常に少ないという現状」はその後も大きく変化していないようである。

こうした現状をふまえ、本稿第一筆者（以下、筆者）は、(1')小学校 1 年生から大学 1 年生までの 13 学年の児童・生徒・学生に、(2'-3')既存の日本語学習者コーパスで使用されたものと同一のイラストを使い、その内容を描写させるストーリーライティング形式で作文を行わせ、(4')集めたデータをダウンロード版・オンライン版として全面的に一般公開する、という基本コンセプトのもと、136,635 語（短単位）からなる新たなコーパスを構築した。本稿は、以下、当該コーパスの構築過程と概要について報告する。

２．JASWRIC の開発経緯

２．１　目的

「小中高大生による日本語絵描写ストーリーライティングコーパス」（Japanese Students' L1 Story Writing Corpus：JASWRIC／ジャスリック）は、2 つの研究目的に基づいて開発された。1 点目は、JASWRIC を単独で使用することで、学齢期の子どもの L1 日本語の書き言葉の発達過程を調査することである。2 点目は、1,000 人の日本語学習者と 50 名の成人日本語母語話者の産出データを収集している I-JAS のストーリーライティングデータと併用することで、L1 の子どもと L1 の成人間の比較や、L1 の子どもと L2 学習者の比較を行うことである。後者について、学習者コーパス研究では、学習者の産出を、参照基準となる母語話者の産出と比較して学習者の過剰・過少使用を特定する中間言語対照分析（contrastive interlanguage analysis）が広く行われているが、少数の母語話者データを参照基準とすることには批判もある。I-JAS の母語話者データを JASWRIC で拡張することにより、参照基準が多様化し、I-JAS を用いた L1/L2 対照分析の妥当性と安定性が向上するという効果も期待される（石川、2022）。

２．２　データ収集の過程

JASWRIC のデータ収集には、関西圏に所在する国立大学法人が設置する 3 つの学校（小学校、中等教育学校、大学）に在籍する児童・生徒・学生が参加した。参加者による作文とアンケート回答データは、以下の 5 期に分けて収集された（表 1）。

表 1 データの収集プロセス

段階	時期	収集したデータ
1 期	2022 年 1～2 月	小 3～小 6 の作文収集
2 期	2022 年 4～5 月	中 1～高 3 の作文・アンケート収集
3 期	2022 年 6 月	小 1～小 2 の作文収集
4 期	2022 年 6～7 月	小 3～小 5（新小 4～小 6）のアンケート収集
5 期	2022 年 7 月	大 1 の作文・アンケート収集

研究倫理審査において承認（次節参照）が得られた時期の関係で、小学校 3～6 年生のデータを先行収集し、以後、中学生・高校生データ、小学校 1～2 年生データ、大学生データの順で収集を進めた。小学校 3～5 年生については、作文収集とアンケート調査を別個に実施したが、そのほかの学年では両者を同時に実施した。

２．３　研究倫理対応

とくに、成年に達していない児童・生徒の作文データを公開目的で収集する際には、研究倫理面でも慎重な対応が求められる。この点をふまえ、小中高生からのデータ収集について

は、前節で示した 1〜4 期のそれぞれにおいて、研究計画（コーパスの構築と公開を目的としていることを明記）を法人の設置する研究内容審査委員会に提出し、「人を直接の対象とする研究の審査」承認を受けた。

また、実際のデータ収集にあたっては、小中高生の保護者向けに、文書でプロジェクトの趣旨を説明し、以下の点を示して同意を求めた。表 2 は第 3 期に使用した依頼フォームの一部である。

表 2　小中高生保護者向け依頼文での説明内容

本調査の参加にかかる確認事項
(1) 本調査への参加は任意です。 (2) 本調査への参加の有無や，提出物の内容が学校での評価などに影響することは一切ありません。 (3) この用紙（裏面）に記名の上、作文を提出いただくことで、本調査への参加と，作文データの公開にご同意いただいたものと取り扱わせていただきます。 (4) 作文を提出した後で、それを取り消したくなった場合は、研究代表者に連絡してください。提出済みの作文を公開データベースから取り除く処理を行います。 (5) 収集した用紙は電子的にスキャンし、書き起こしを済ませた後、適切な方法で廃棄します。作文本文は電子的データベースとして公開されます。

大学生については、全員が成年（18 歳）に達していることから、上記と同様の趣旨を文書で学生本人にのみ示し、同意を求めた。

小中高生の参加者には一切の報酬を与えておらず、ボランティアとしての参加となっている。そのため、学年間で参加者数に差が生じている。一方、大学生参加者には、仕事の内容と少額の報酬を提示した上で参加者を募った。

２．４　参加者

参加者の内訳は以下のとおりである。中等教育学校において生徒の学年は 1〜6 年と呼称されているが、以下では、一般的な校種区分に従い、中学校 1〜3（中 1〜中 3）年および高等学校 1〜3 年（高 1〜高 3）と記載している。また、データ整理の必要上、すべての学年（grade : G）を通して G01〜G13 という学年コードを設定した。なお、ストーリーライティングは後述のように 2 つの課題からなるが、一方しか回答していないデータはコーパスから削除している。これにより、最終的な参加者人数は 700 名となった（表 3）。

表 3 JASWRIC 参加者人数

学年	人数	作文数	学年	人数	作文数
G01（小 1）	37	74	G08（中 2）	112	224
G02（小 2）	43	86	G09（中 3）	45	90
G03（小 3）	35	70	G10（高 1）	90	180
G04（小 4）	27	54	G11（高 2）	86	172
G05（小 5）	58	116	G12（高 3）	30	60
G06（小 6）	62	124	G13（大 1）	53	106
G07（中 1）	22	44	以上合計	700	1,400

前述のように、小学校 3〜6 年生の作文は 2022 年 1〜2 月に先行収集され、そのほかの参加者の作文は、2022 年の 4〜7 月に収集された。このため、小学校 3〜6 年生については当該年度の後半で、そのほかの学年では当該年度の前半でデータが収集されたことになる。年度内での執筆時期の数カ月のずれが産出に決定的な影響を与えることは少ないと思われる

が、13 年間の発達過程を精密な計量的モデルとして議論する場合には、たとえば、G1、G2、G3.5、G4.5、G5,5、G6.5、G7、G8…のように、小学校 3〜6 年生のデータのみ年次数を割り増して分析することも考えられるだろう。

２．５　作文の収集

作文コーパスでは、様々なタイプの作文が収集対象にされるが、今回のプロジェクトでは、2.1 節で述べた目的をふまえ、I-JAS で実施された 2 種のストーリーライティングタスク(「鍵」をテーマにした 4 枚の連続イラストに基づく作文と、「ピクニック」をテーマにした 5 枚の連続イラストに基づく作文）を実施することとした。自由作文に比べ、絵描写作文は内容的なぶれがきわめて小さく、対照研究に適した資料と言える。I-JAS の 2 種のイラストの使用については、コーパス開発者の迫田久美子氏より事前に許諾を得た。

I-JAS では、同じイラストについて、先にストーリーテリング（口頭描写）を行ってからストーリーライティング（コンピュータ上での作文）を行う手順になっているが、JASWRIC では、作文タスクのみを実施している。

I-JAS のオリジナルのイラストでは、キーワードが日本語・平仮名・英訳で示されている。「鍵」については、鍵（かぎ／key)、ケン（けん)、マリ（まり)、梯子（はしご／ladder)、警官（けいかん／police man）の 5 語が、「ピクニック」については、ピクニック（ぴくにっく／picnic)、ケン（けん)、マリ（まり)、バスケット（ばすけっと／basket)、犬（いぬ／dog)、地図（ちず／map）の 6 語がキーワードである。今回のプロジェクトでは、英訳はすべて除去した。そのうえで、小学校 1〜2 年生用のプロンプトでは、キーワードの漢字は「犬」を除いてすべて平仮名で表記し、「犬」は漢字に平仮名を添え、カタカナ語はカタカナに平仮名を添えた。小学校 3 年生〜大学 1 年生用のプロンプトでは、英訳を除いたほかは、オリジナルのキーワード表記をそのまま使用した。

以下の図 1–2 は、小学校 3 年生以上の参加者に配布したプロンプトである。I-JAS のオリジナルの調査デザインに合わせ、書き出しの指定文（漢字使用は学年によって調整）を作文欄の冒頭に明示し、続きを書かせるようにした。

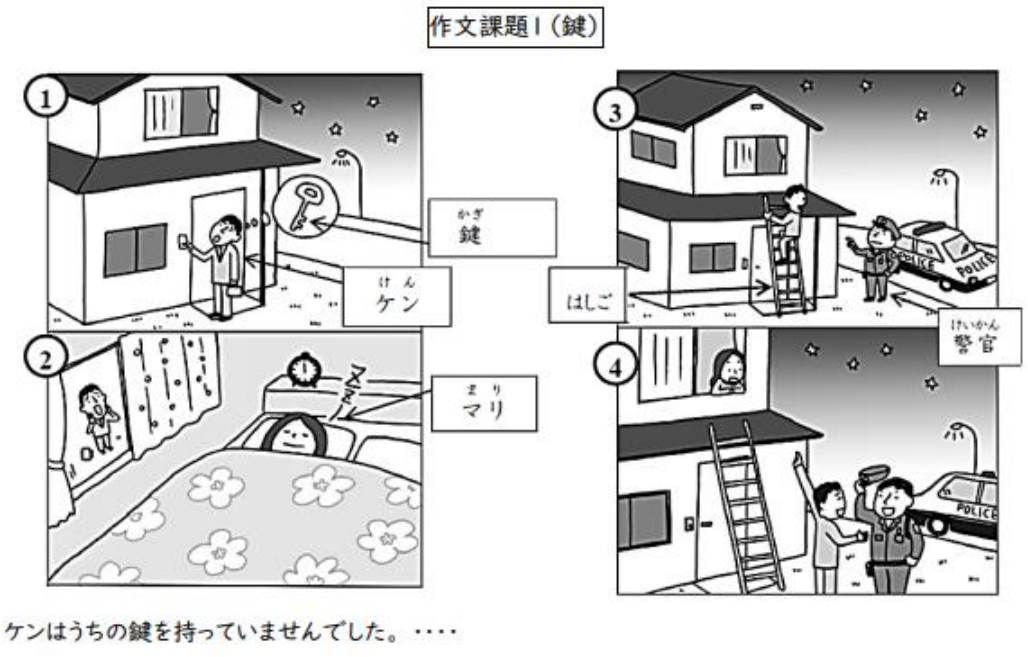


図 1 ストーリーライティング用のプロンプト（「鍵」）

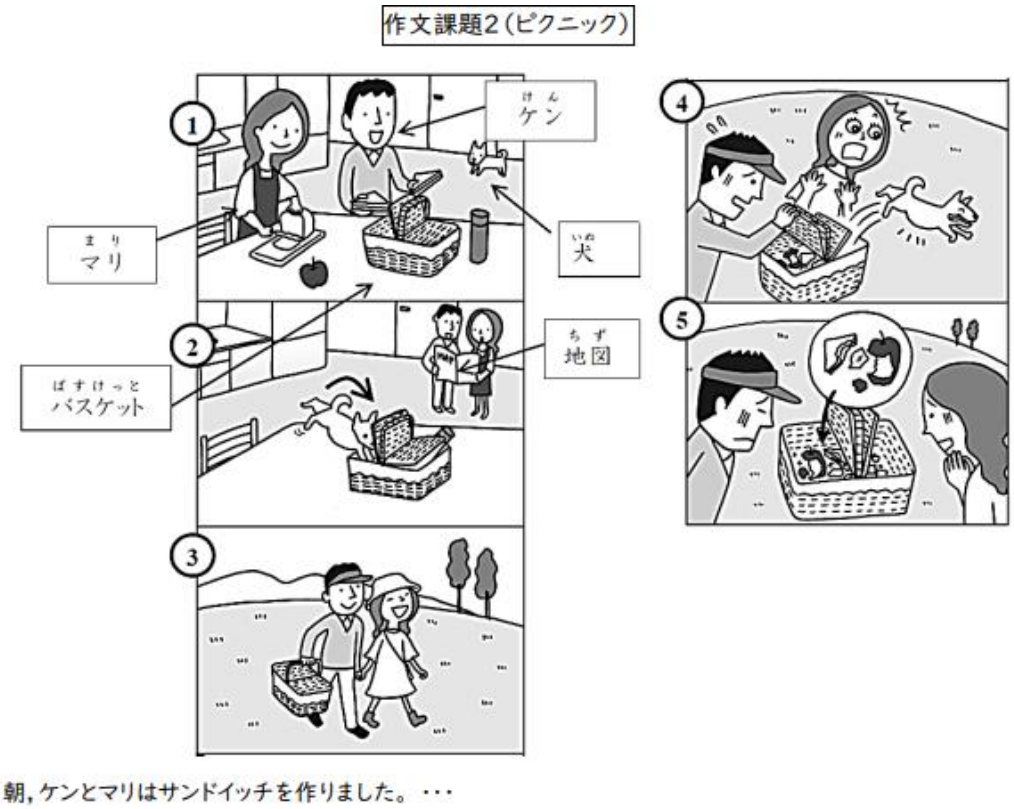


図 2 ストーリーライティング用のプロンプト（「ピクニック」）

以下は、学年グループ別の課題指示文である。小学校 3 年生～高校 3 年生用では、すべての漢字にフリガナを付けている（表 4）。

表 4　参加者向け課題指示文

学年	指示文
小 1～2	えをよくみてください。（1～4／1～5 のじゅんばんにならんでいます）。かきだしのことばにつづくよう、4 つ（5 つ）のえがあらわすストーリー（おはなし）をかいてください。
小 3～高 3	絵［え］をよく見［み］てください。 書［か］き出［だ］しのことばに続［つづ］くよう、4 つ（5 つ）の絵［え］が表［あらわ］すストーリー（お話「はな」し）を書［か］いてください。
大 1	下記の 2 種の絵をよく見てください。書き出しの言葉に続くよう，①～④（または①～⑤）の絵が表すストーリー（お話）を書いてください。

執筆方法に関して、小学校 1 年生～高校 3 年生では、教員を通じて用紙を配布し、用紙内の指定箇所に筆記具で記入するよう指示した。大学生は、プロンプトを電子ファイルで配布し、作文はオンラインのサイト上に記入するよう指示した。いずれの場合も、字数制限と時間制限は設けず、平仮名の使用は問題ないと示した。また、参加者本人が独力で作文するよう示した。小中高生調査では、児童生徒への指示に加え、保護者向けの依頼文中にも注意事項を明記した（表 5）。

表 5　小中高生保護者向け依頼文での課題指示（小 1〜2 保護者用）

保護者用の課題指示
作文の執筆にあたって、習っていない漢字はひらがなで書いていただいて問題ございません。時間制限は設けておりませんが、所要時間はおよそ 5〜10 分程度と予想されます。なお、学年間の比較が調査の趣旨でありますことから、児童ご本人以外の方のお手伝いはお控えいただきまして、ご本人が独力で書いてくださるようお願いを申し上げます。

２．６　アンケートデータの収集

参加者の属性データとして、小学校 6 年生を除く全参加者に対して、「本を読むことが好きですか？」および「文を書くことが好きですか？」という 2 つの質問を与え（質問文中の漢字使用は学年によって調整している）、5 段階での回答データを収集した。小中高生については、客観的な学力データの収集を行っていないが、2 つの質問によって、読書や作文へのなじみ度は推定可能である。

小学校 1〜2 年生および中学生〜大学生においては、2 項目のアンケートを作文用紙に記載し、作文と同時に提出させた。ただし、前述のように、先行してデータ収集を行った小学校 3〜6 年生については、作文時にアンケートを実施していなかったため、作文収集の 4〜5 カ月後に、アンケートのみを別個に実施した。アンケート実施時点で学年が進行していたため、一定数が法人設置学校以外に進学した 6 年生についてはアンケートデータが取られていない。

また、大学生の参加者には、上記のアンケートに加え、2022 年 1 月に実施された大学入学共通テストにおける国語（現代文）のスコアレンジ（100 点満点中 10 点刻み）を申告させた。参加者の性別は、産出への影響もある程度予想されるところであるが、性多様性への配慮の観点からあえて尋ねていない。

２．７　作文の文字化

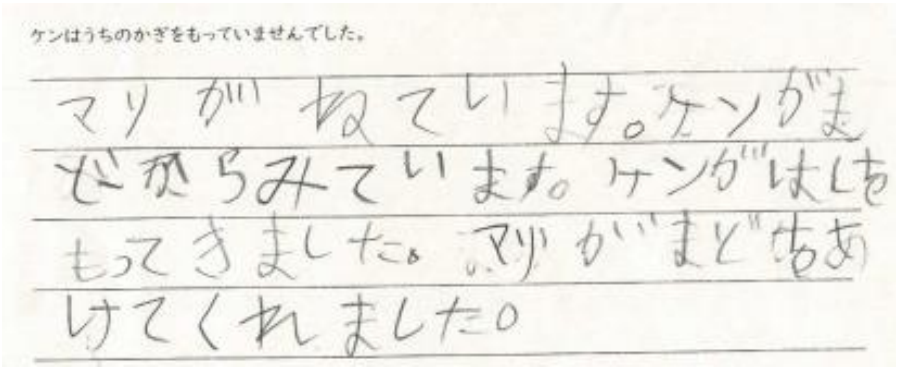
小中高生の作文は手書きでなされたため、作文用紙をスキャンして専門業者に回送し、書きおこしを依頼した。書きおこしの質にぶれが出ないよう、作業者は 20 年以上の作業経験を有する専業者 1 名のみに限定した。書きおこしにあたっては、以下のルールを定め、作業者に指示した（表 6）。

表 6 書きおこし作業マニュアル

書きおこし作業指示（中高生データ処理の指示書の一部）
(1) 入力範囲 ・生徒が手書きで書いた作文本体（※欄外のメモや、注記などはカット：原則として青でマーク済） ・書き出しの印刷された指示文は入力不要（ケンはうちのカギを持っていませんでした／朝、ケンとマリはサンドイッチを作りました） ・コマ番号を示す①、②などの番号を生徒が入れている場合、それらはカットして入力 (2) 生徒による修正の扱い ・二重線や塗りつぶしではっきり消しているものは消されたものとみなして入力しない ・消して書き直しているものは、書き直したほうを入力する ・一度文を書いてから、山形マークなどで別要素を挿入しているものは挿入されたものとして入力。山形記号の反映は不要　私は＜後から追記：イヌと＞遊びに行った　→　入力例：　私はイヌと遊びに行った (3) 誤字・読みにくい文字の扱い ・軽微な誤字や悪筆（点が足りない、線が一本多いなど）は、常識的に読み替えて入力（貝る→見る）

・明確な誤字は【誤】タグをつけてそのまま入力、（大無し→【誤】大無し）
(4) 記号・特殊表記
・記号もそのまま入力（！？など、すべて全角で。！については直線・斜線・太い細いなどの区別があるが一律で！で OK）
・「おま w」　（おまえ、の意味のネット用語）などもそのまま入力
・本来は「　」が必要なセリフに、「　」記号がない場合が散見されるが、ママで入力
(5) 段落扱い
・明確な改行がない場合は続けて入力
・明確に行替えがある場合は改段として処理

以下は、小学校 1 年生の児童作文の現物と、入力後のデータの一例である（図 3）。



マリがねています。ケンがまどからみています。ケンがはしをもってきました。マリがまどをあけてくれました。

図 3 小学校 1 年生の「鍵」作文の児童手書き原稿と書きおこしデータ（G01_Key_009）

２．８　データファイルの作成

個々の作文ファイルは UTF-8（BOM あり）形式で保存し、以下の基準で命名を行った（表 7）。

表 7 ファイル命名ルールと実際の例

凡例	例
学年コード_トピックコード_参加者連番.txt	(1) G01_Key_032.txt（小 1、032 児童の「鍵」作文）) (2) G10_Pic_006.txt（高 1、006 番生徒の「ピクニック」作文）

２．９　テキスト校閲と形態素解析

形態素解析には、国立国語研究所が提供する「Web 茶まめ」を使用した。前処理として、「半角・全角変換」と「数字処理」オプションを追加し、辞書は現代語を使用した。当初、素起こしのテキストデータに対して形態素解析を試みたが、とくに小学生低学年の児童はほぼすべてを平仮名で表記しており、かつ、句読点を使うことが少ないため、解析精度は著しく低かった。

そこで、解析の精度を上げるため、全データに対して、筆者が手作業で校閲を加えた。以下は、実際に行った校閲のタイプ別事例である（表 8）。

表 8 元テキストに対する校閲の例

校閲タイプ	元テキスト	校閲後テキスト
(1) 平仮名の漢字化	・ねむってたらだれかがよんできた ・通ほうされて… ・と中 ・じゅう人	眠ってたら誰かが呼んできた ・通報されて… ・途中 ・住人
(2) 平仮名のカタカナ化	・ちゃいむをならしました… ・ぴくにっク ・まりはねてたからけんは… ・ぐーすかぴーすかねています	・チャイムをならしました… ・ピクニック ・マリは寝てたからケンは… ・グースカピースカ寝ています
(3) 文節区切りスペース挿入	・さんどいっちがありませんでしたおなかぺこぺこで…	・サンドイッチがありませんでした［スペース］お腹がペコペコで…
(4) エラー修正	・きずきません ・事【誤】状を説明すると…	・気づきません ・事情を説明すると…
(5) 段落削除	・バスケットの中から犬がとび出してきました。【改段】マリはとてもおどろきました。【改段】2 人がバスケットの中を見ると…	・バスケットの中から犬がとび出してきました。マリはとてもおどろきました。2 人がバスケットの中を見ると…

(1) については、漢字で書くのがふつうであると判断した箇所と、平仮名で書いた場合に異なる解釈の余地が残ると判断した箇所（あける→開ける／空ける）を対象とした。とくに、「通ほう」「と中」「じゅう人」など、部分的に漢字表記されているものについては、そのままでは解析が失敗することが多いため、全体を漢字で表記するようにした。

(2)については、カタカナで書くのがふつうであると判断した箇所のほか、特殊なオノマトペと判断されるものについても地の文との切れ目を示すため、カタカナ表記に変更した。

(3)については、使用した句読点の数などを調査対象にする研究が想定されることから、句読点を校閲で追加することは行わず、スペースを補って、語の切れ目を示した。なお、補ったスペースは形態素解析後、再度消去している。

(4)については、書きおこし作業者が【誤】タグを付与した箇所のほか、MS Word の校閲でエラーと指摘された箇所を中心に修正を行った。複数の参加者に共通して見られたエラーとしては、以下のようなものがある（表 9）。

表 9 参加者による特徴的なエラー

タイプ	元テキスト	校閲後テキスト
(1) 「は」／「わ」の混同	・けんわ	・ケンは
(2) 「を」／「お」の混同	・はしごお	・はしごを
(3) 促音の不足･過剰挿入･挿入場所誤り	・のぼて ・さんどっいっち ・はっいた	・のぼって（登って） ・サンドイッチ ・はいった（入った）
(4) 四つ仮名の誤り	・きずいて	・気づいて
(5) 漢字の誤り	・事状 ・不信（者） ・窓が空いている	・事情 ・不審（者） ・窓が開いている

筆者による以上の校閲は、後で検証・追加できるよう、すべて、Word の校閲機能を用い

て記録した（図 4）。

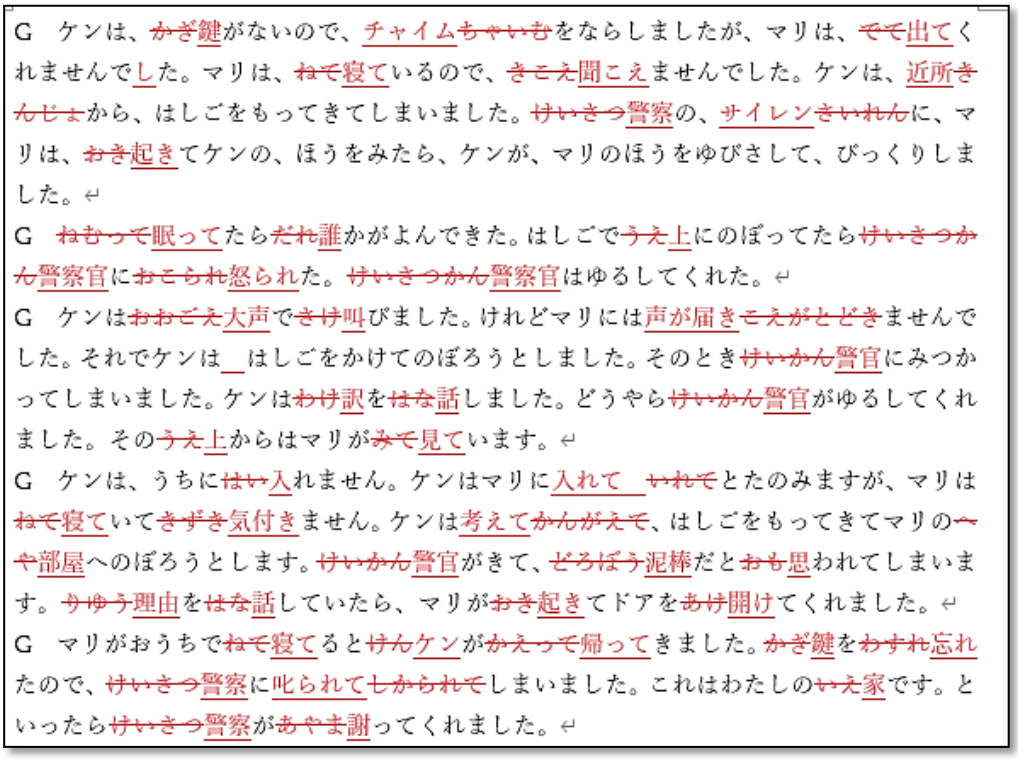

G　ケンは、~~かぎ~~鍵がないので、チャイム~~ちゃいむ~~をならしましたが、マリは、~~そで~~出てくれませんでした。マリは、~~ねて~~寝ているので、~~きこえ~~聞こえませんでした。ケンは、近所~~きんじょ~~から、はしごをもってきてしまいました。~~けいさつ~~警察の、サイレン~~さいれん~~に、マリは、~~おき~~起きてケンの、ほうをみたら、ケンが、マリのほうをゆびさして、びっくりしました。↵
G　~~ねおって~~眠ってたら~~だれ~~誰かがよんできた。はしごで~~うえ~~上にのぼってたら~~けいさつかん~~警察官に~~おこられ~~怒られた。~~けいさつかん~~警察官はゆるしてくれた。↵
G　ケンは~~おおごえ~~大声で~~さけ~~叫びました。けれどマリには声が届き~~こえがとどき~~ませんでした。それでケンは＿はしごをかけてのぼろうとしました。そのとき~~けいかん~~警官にみつかってしまいました。ケンは~~わけ~~訳を~~はな~~話しました。どうやら~~けいかん~~警官がゆるしてくれました。その~~うえ~~上からはマリが~~みて~~見ています。↵
G　ケンは、うちに~~はい~~入れません。ケンはマリに入れて　~~いれて~~とたのみますが、マリは~~ねて~~寝ていて~~きずき~~気付きません。ケンは考えて~~かんがえて~~、はしごをもってきてマリの~~へや~~部屋へのぼろうとします。~~けいかん~~警官がきて、~~どろぼう~~泥棒だと~~おも~~思われてしまいます。~~りゆう~~理由を~~はな~~話していたら、マリが~~おき~~起きてドアを~~あけ~~開けてくれました。↵
G　マリがおうちで~~ねて~~寝てると~~けん~~ケンが~~かえって~~帰ってきました。~~かぎ~~鍵を~~わすれ~~忘れたので、~~けいさつ~~警察に叱られて~~しかられて~~しまいました。これはわたしの~~うえ~~家です。といったら~~けいさつ~~警察が~~あやま~~謝ってくれました。↵

図 4 筆者による元データへの校閲追加の例

全 1,400 作文に対する校閲箇所は、延べで 9,090 箇所であった（4,631 箇所の挿入と、4,459 箇所の削除）。これらの校閲は、筆者が目視で気づいた範囲で、かつ、解析精度を高めるという目的に即して行ったもので、参加者の作文に見られる問題点を網羅的に検出・修正する性質のものではない。

校閲により、解析精度は、完全ではないものの、かなり上昇した。たとえば、上記にある「ちゃいむ」「さいれん」「けいさつかん」の部分について言うと、平仮名表記の元テキストではそれぞれ「ちゃう／無」「然（さ）／入れる」「警察／館」と誤解析されていたが、校閲テキストではいずれも正しい解析がなされている。V1.0 の公開に先立ち、人名である「ケン」と「マリ」の品詞情報の不統一のみ修正したが、そのほかは、自動解析結果のままで、手作業による修正は加えていない。

２．１０　データの概要

以上の手順を経て、公開用のデータが完成した。表 10 は JASWRIC（V1.0）に含まれる作文の総語数（短単位解析による形態素の数）である。

「鍵」作文の総語数は約 72,000 語、「ピクニック」作文の総語数は約 65,000 語、合計約 136,000 語となる。近年の大規模なコーパスから見れば、必ずしも大きな数字ではないが、統制性の高い共通の絵描写ストーリーライティングであることを考えれば、個体差を超えた全体傾向を議論するに足る量のデータが集められたと言える。なお、今後、形態素解析結果の修正がなされた場合、語数は若干増減する可能性がある。

表 10 JASWRIC V1.0 の収録語数データ

Grade	人数	総語数			1 人あたり語数		
		Key	Pic	All	Key	Pic	All
G01	37	2328	2446	4774	62.9	66.1	129
G02	43	3084	2683	5767	71.7	62.4	134.1
G03	35	3884	3461	7345	111	98.9	209.9
G04	27	3023	2742	5765	112	101.6	213.5
G05	58	6045	5611	11656	104.2	96.7	201
G06	62	6080	5451	11531	98.1	87.9	186
G07	22	2839	2158	4997	129	98.1	227.1
G08	112	10499	9597	20096	93.7	85.7	179.4
G09	45	4623	4323	8946	102.7	96.1	198.8
G10	90	10811	9343	20154	120.1	103.8	223.9
G11	86	9597	8561	18158	111.6	99.5	211.1
G12	30	3545	2980	6525	118.2	99.3	217.5
G13	53	5700	5221	10921	107.5	98.5	206.1
All	700	72058	64577	136635	102.9	92.3	195.2

３．JASWRIC の公開

３.１　ダウンロード版

JASWRIC（V1.0）は、2022 年 8 月よりダウンロード版としての公開を開始した。ダウンロードデータには、Raw Data と Edited Data の 2 つのサブフォルダが含まれる。

前者には、個別作文の素起こしテキストファイル（txt）1,400 種、学年別・トピック別にテキストファイルを集約したマージテキストファイル（txt）26 種、また、小中高生 647 名による手書き作文のスキャン画像（jpg）1,294 枚（図 5）が含まれる。スキャン画像のデータは、子どもの文字の獲得研究などにも応用可能であろう。

後者には、形態素解析用に校閲を加えた文書ファイル（docx）が含まれる。文書ファイルには、Word の校閲機能を用い、素起こしのテキストに加えた編集履歴がすべて記録されている。

このほか、形態素解析済みデータ（tagged）、参加者属性シート（Participant Survey）、参考文書（readme）が同梱されている。

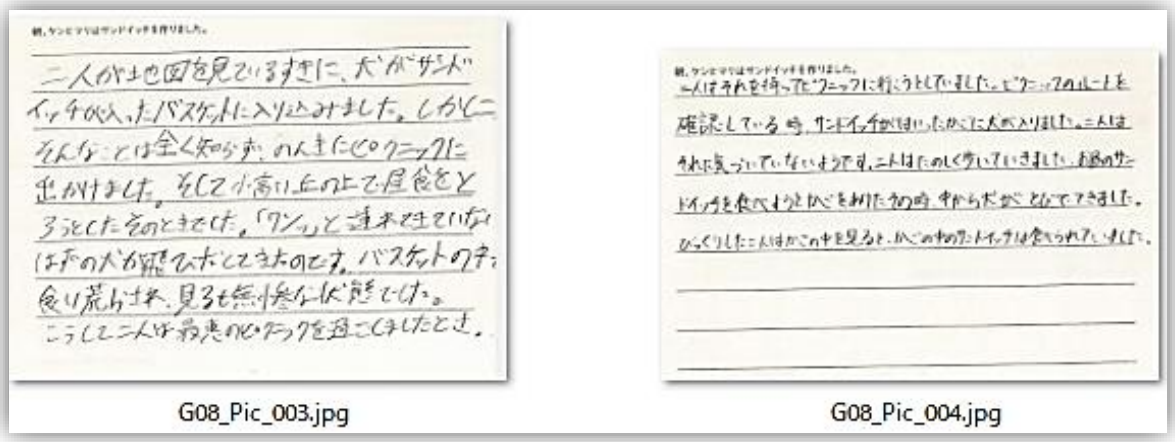


図 5 手書き作文の画像ファイル（中 2 生徒の作文の一部）

３．２　オンライン検索版

３．２．１　概要

JASWRIC は、筆者の研究室でこれまでに開発した各種コーパスのオンライン統合検索サイト（Kobe Univ. Ishikawa Lab. Online Corpus Retrieval System）上でも公開されている（図 6）。

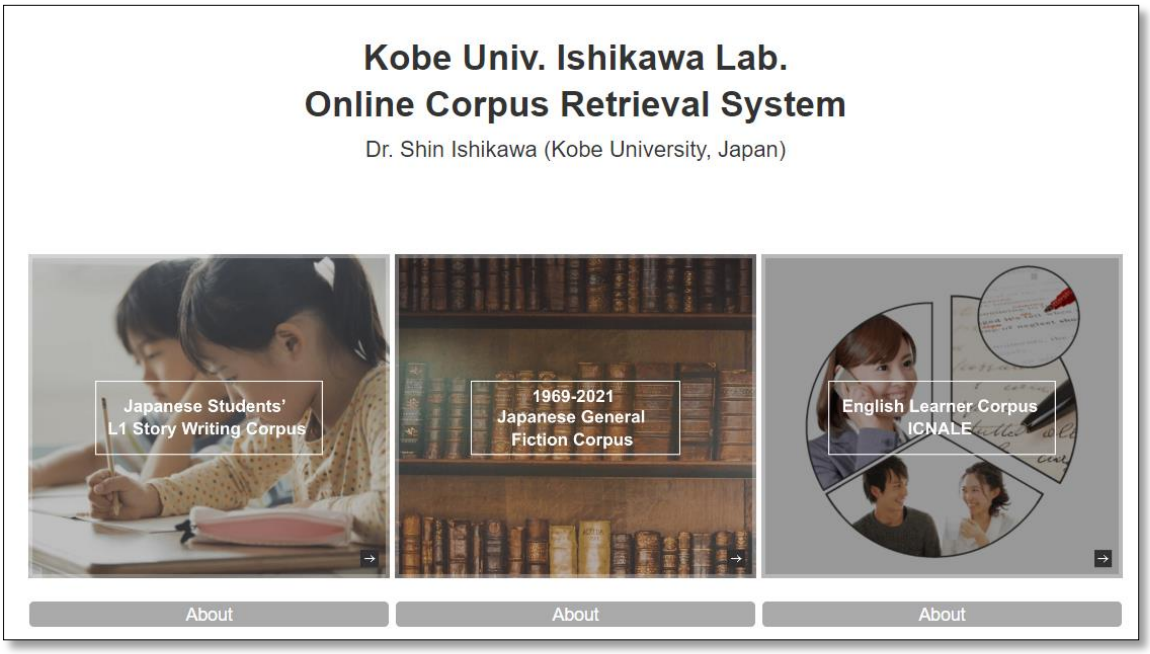


図 6 Kobe Univ. Ishikawa Lab. Online Corpus Retrieval System トップページ
（左から順に、JASWRIC、日本語小説コーパス JFIC、英語学習者コーパス ICNALE）

留意すべきは、オンライン版は、語彙研究や品詞研究などでの利用を念頭に置いているため、校閲作業後のデータを使っていることである。つまり、低学年児童の作文中の平仮名などは適宜漢字に直されており、このほか、大きなエラーなども修正されている。このため、漢字使用率や、誤字・誤記などを主たる関心対象として研究を行いたい場合は、前述のダウンロード版に含まれる生データを、直接、分析するほうが便利であろう。ただし、後述するように、オンライン版であっても、検索結果から、手書き作文の画像ファイルへのアクセスが提供される設計となっている。

オンライン版（JASWRIC *Online*）では、子どもの日本語作文の多角的な分析が可能になるよう、既存コーパス用に研究室で開発した検索システムを内部的に改修して運用している。基本的な検索タイプは共通化されており、(1)KWIC 検索（検索対象語を中央に、文脈を左右に配置して一覧表示）、(2)共起語検索（当該語の左右 3 語範囲で位置ごとの高頻度共起語を表示、統計値は頻度・*t* スコア・対数尤度比・相互情報量から選択可能）、(3)語彙リスト作成（書字形［=表層形］または語彙素に基づく語彙頻度表を作成・出力）、(4)頻度グラフ作成（検索対象語の学年別頻度変化をグラフで表示、頻度は総語数で自動的に調整頻度で出力）、(5)特徴語検索（ターゲットデータとレファレンスデータに内包されるすべての語の頻度を比較して、ターゲット側で有意に多い、または少ない語をキーワードとして一覧で表示。統計値はカイ二乗値と対数尤度比から選択可）の 5 種の機能を提供する。

図 6 の左端の写真をクリックすると、ユーザー登録の画面に遷移し、必要事項を記入すると、図 7 に示す、KWIC 検索用の検索条件指定画面が表示される（以下、本稿の画像は開発版のシステムから取ったもので、運用版では若干の変更が生じる可能性がある。また、実際の画面の一部を加工して表示している場合がある）。

図 7 KWIC 検索の検索条件指定画面（一部）

上部にあるタブを押すことで、5 つの検索機能を切り替えることが可能で、選択中の機能は赤で表示される。KWIC 検索の条件設定画面について言うと、まず、1 行目にある「検索語」に調べたい語（または語の連鎖）を入力する。JASWRIC *Online* では、上記の「検索語」を含め、コーパス分析の基本概念や基本技術に関わる用語にはリンクが用意されており、リンクを押すことで、解説などが表示される。次に、2 行目の「学年」で、調査したい学年を設定する。上部にある「小学校」「中学校」「高校」「大学」をクリックすると、それぞれ、自動的に、下部にある全学年が指定されるが、必要に応じてその一部を検索対象外にすることもできる（たとえば、小学校の低学年にあたる 1〜2 年生のみを調べるなど）。最後に、3 行目の「トピック」で、ストーリーライティングにおける 2 種類の課題のうち、調査対象としたいものを、一方または両方とも、指定する。

３．２．２　検索語の入力方法

どの検索を行う場合であっても、検索語の入力方法は同一であるが、ここでは、KWIC 検索に即して、検索語の入力の流れを確認しておきたい。

まず、検索は、書字形の単位で行われる。ゆえに、検索ボックスに「犬」と入れれば「犬」を含む用例が、「する」を入れれば「する」を含む用例が抽出される。ただし、語彙素をまとめているわけではないので、「する」を入れても、その活用形（「さ（れる）「し（ない）」「すれ（ば）」など）は抽出されない。

活用形を同時に検索したい場合は、縦棒（|）を使って OR 検索を行う。たとえば、「さ|し|する|すれ|しろ」と入力すれば、「する」とその活用形を同時に検索することができる。

JASWRIC *Online* は、書字形を、形態素（短単位）単位で区切ったデータを元データとしているため、連語や複合形を検索する際は、形態素の切れ目ごとに、半角のスペース（以下では[sp]で表示）を挟む必要がある。たとえば、「見ている」という表現を探すには、「見[sp]て[sp]いる」と入力し、「見ていた」を探すには「見[sp]て[sp]い[sp]た」と入力する。、自分の探したい表現がどのように区切られるかわからない場合は、「Web 茶まめ」（2.9 節参照）に当該表現を入力し、どのように区切られるか確認すると便利だろう。

品詞情報を検索条件に加えることもできる。検索ボックスの右側にある[POS]を押すと、

以下の品詞選択画面がポップアップ表示される（図 8）。

図 8 品詞指定画面（大区分に名詞を選んだ場合）

品詞選択画面は上段（大区分）と下段（小区分）の二階層構造になっており、上部（赤色で表示）で品詞の大区分を指定すると、下部（グレーで表示）に当該品詞の小区分が表示される仕様となっている。下段に表示されるボタンのいずれかを押すと、当該品詞を示すコードが自動的に検索ボックスに入力される。なお、[ALL]とは全小区分の同時指定のことである。

たとえば、上図で示されている名詞を例にして言えば、小区分のうち、固有名詞を選ぶと絵課題の登場人物名である「ケン」や「マリ」が、普通名詞を選べば「鍵」や「チャイム」が、数詞を選べば「二（階）」などが抽出される。品詞と書字形を連語として指定することもできる。たとえば、「[名詞] は」と入れると、名詞の後に「は」が後続する例が抽出される。

検索語の入力にあたって、注意すべき点は、異表記が別語扱いになることである。たとえば、「鍵」と「カギ」は表記が異なるため、どちらか一方を入力した場合、他方は出てこない。両方を検索するには、OR 検索を用いて「鍵|カギ」のように入力する必要がある。なお、OR 検索は、連語の一部に対しても使える。たとえば、「ケン|マリ[sp]は|が|の|を|に」と入力すれば、「ケンは」「マリが」「ケンの」など、「ケン」または「マリ」の直後に主要な助詞が後続する例を同時に検索することができる。

以下は、主な検索ルールを一覧にしたものである（表 11）。

表 11 JASWRIC *Online* 上での検索指定の例

検索タイプ	検索指定の例	出力例
通常	鍵	鍵
OR 検索指定	鍵|カギ	鍵、カギ
連語・複合形指定	し[sp]て[sp]いる	…している
品詞指定（大区分）	[名詞]	鍵、チャイム、梯子…
品詞指定（小区分）	[名詞-固有名詞]	ケン、マリ、日本、ポチ…
品詞＋書字形指定	[名詞] [sp]は	鍵は、チャイムは、梯子は…
OR 検索組み合わせ指定	ケン|マリ[sp]は|が|	ケンは、ケンが、マリは、マリが

コーパス分析に詳しくないユーザーの場合、検索語の入力ルールはやや複雑に思えるかもしれないが、検索ボックス右側の[?]ボタンを押せば、用例付きの解説をオンラインで読むことができる。なお、本システムの検索精度は、データベースの元になっている形態素解析の精度に制約されており、完全なものではない。

３．２．３　KWIC 検索

以下、JASWRIC *Online* で提供されている各種の検索機能について概説する。まず、KWIC 検索は、関心対象語の文脈内での振る舞いを調べるために用いられる。図 9 は、全データを対象に「入っ[sp]た」を KWIC 検索した際の結果画面（一部、以下同）である。

図 9 KWIC 検索の結果画面の一部（検索語：「入っ た」）

これらの各行をコンコーダンスラインと呼ぶ。コンコーダンスラインの中央には検索対象語が表示され、左右に一定の長さの文脈が表示される。無作為な順序で並んだコンコーダンスラインから傾向を読み取ることはむつかしいが、たとえば、「入った」の左隣ないし右隣に来る共起語の 50 音順で全体の行を並べ替えると（ソーティング）、同じ共起語が続けて並ぶこととなり、全体を概観することで、検索対象語がどのような語と一緒に生起しやすいかを直観的に読み取ることができる。

ソーティングの基準は、上記の画面上部の 3 つのキーで指定する。たとえば、第 1 キーを L1（左隣 1 語目）、第 2 キーを L2（同 2 語目）、第 3 キーを L2（同 3 語目）と指定すれば、「入った」の左隣 1〜3 語目を基準として並べ替えが行われる。

また、各行の右端には参加者コードが表示されている。たとえば、上記の 1 行目の G03_012_Pic は、小学校 3 年生（G03）、学年内コード 12 番の児童で、これがピクニック（Pic）作文の一部であることがわかる。ダウンロード版に含まれる参加者メタデータを参照すれば、当該の児童が、読書と作文に対してそれぞれどの程度の親しみを感じているか、アンケート結果を確認することができる。

各行の左端にあるアイコンは、手書き作文の画像ファイルへのリンクを示す。たとえば、1 行目左端のアイコンを押すと、以下の画面がポップアップ表示される（図 10）。

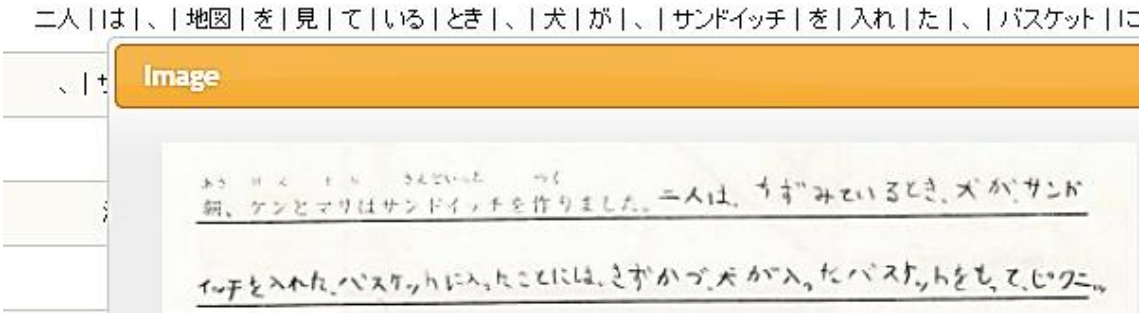


図 10 KWIC 検索結果画面で手書き画像を表示させた場合の例

こうした手書きデータを見ることで、実際に参加者が書いた作文を確認しながら分析を進めることができる。なお、大学生についてはワープロ上で作文したため、手書き画像データへのリンクはない。

３．２．４　共起語検索

共起語検索とは、検索対象語が出現しやすい近傍部の語彙的環境を明らかにする目的でなされる。以下は、全データを対象に「する」の共起語を検索した際の結果画面である（図11）。

Raw Freguency　t-score　Log-Likelihood　Mutual Information

	L1		0	R1		R2		R3	
78 (5484)	と	90 (4327)	する 475	と	382 (4327)	警官	53 (930)	が	55 (46
26 (322)	説明	55 (240)		場所	17 (141)	に	31 (5913)	警官	48 (9
25 (276)	を	36 (5484)		こと	15 (701)	マリ	15 (2596)	は	27 (41
23 (810)	確認	28 (207)		ため	12 (231)	警察	13 (226)	に	23 (59
15 (5913)	そう	8 (206)		公園	5 (111)	犬	10 (1596)	マリ	17 (25
11 (8117)	ピクニック	6 (810)		の	4 (4061)	中	9 (1205)	犬	17 (15
8 (105)	た	5 (8117)		ケン	3 (2132)	を	9 (5484)	から	14 (10

図 11　共起語分析結果画面の一部（検索語：「する」）

検索結果の上部には、4 種の統計量の切り替えボタンがあり、選択中の統計量が赤色で表示されている。デフォルトでは、粗頻度（raw frequency）の降順で各位置における高頻度共起語が示される。上記を見ると、左側要素との組み合わせでは「～とする」「～をする」「そうする」のほか、「説明する」「確認する」のような漢語動名詞が多いことがわかる。また、右側要素との組み合わせとしては、接続詞となる「すると」が最多で、そのほか、「～する場所」「～する公園」のような名詞修飾、また、「～すること」「～するため」といった抽象的名詞を含む連体表現が多いことも確認できる。

粗頻度は、共起を観察する際の一般的な基準だが、必要に応じて、*t* 統計量、対数尤度比、相互情報量に切り替えて結果を比較することができる。これらはいずれも共起の強さを示す尺度で、共起形の頻度を分子に、共起を構成する各要素の単独頻度を分母に置いて計算されるが、計算の詳細には差がある（石川、2021）。3 種の統計量のうち、相互情報量（mutual information score）は、頻度の情報を対数化して圧縮するため、絶対的な頻度は低くても結びつきの強い組み合わせが上位に評価されやすい。たとえば、上記の条件で、統計量を相互情報量に切り替えると、L1 位置では、「プレゼント」などが上位に来る。この場合、「プレゼントする」という共起形の頻度は 1 例のみであるが、「プレゼント」の単独頻度も 1 例である。つまり、「プレゼント」という語はそもそも出現すること自体が極めてまれであるのに、出現する場合には必ず「する」を伴うものとして、上位に評価される。相互情報量は低頻度だがユニークな語の組み合わせを探したい時に有用である。

３．２．５　語彙リスト作成

語彙リスト作成は、指定した参加者群別、トピック別に、使用された語と頻度の一覧を調べる目的でなされる。この時、語の単位として、書字形または語彙素のいずれかを選ぶことができる。書字形を選べば表記や活用形は区別され、語彙素を選べばそれらは集約される。以下は、トピックを「鍵」に限定し、書字形単位でリスト作成を行った際の結果画面である（図 12）。

Word ▲ | Word ▼ | Raw Frequency ▲ | Raw Frequency ▼

Word	Raw Frequency	PMW
て	4381	682.43
た	4185	651.89
を	2750	428.36
に	2706	421.51
まし	2401	374.00

図 12 語彙リストの一部（トピック：「鍵」、書字形単位）

結果画面の上部には、4 つのボタンがあり、デフォルトでは粗頻度（raw frequency）による降順（▼）が選ばれているが、目的に応じて、昇順（▲）に切り替えたり、あるいは、語の 50 音順（word）による昇順や降順に切り替えたりすることも可能である。

リストを見ると、接続助詞の「て」や、過去助動詞の「た」、格助詞の「を」「に」、丁寧形助動詞「ます」の活用形である「まし」などがとくに多いことがわかる。個々の頻度は粗頻度のほか、100 万語あたりの調整頻度（PMW）でも表示される。たとえば学年間で同一語の頻度を比較するといった目的であれば、調整頻度の参照が必要になろう。

同じデータに対して、語彙素を単位として語彙リストを作成すると、3 位に「ます」が入る。これは、書字形で 5 位であった「まし」の頻度に、「ませ（ん）」や「ます」の頻度が繰り込まれるためである。

３．２．６　頻度グラフ作成

頻度グラフ作成は、検索対象語の出現頻度の学年間変化を観察する目的でなされる。以下は、全データを対象として、「そして」の学年別頻度を調査した際の結果画面である（図 13）。

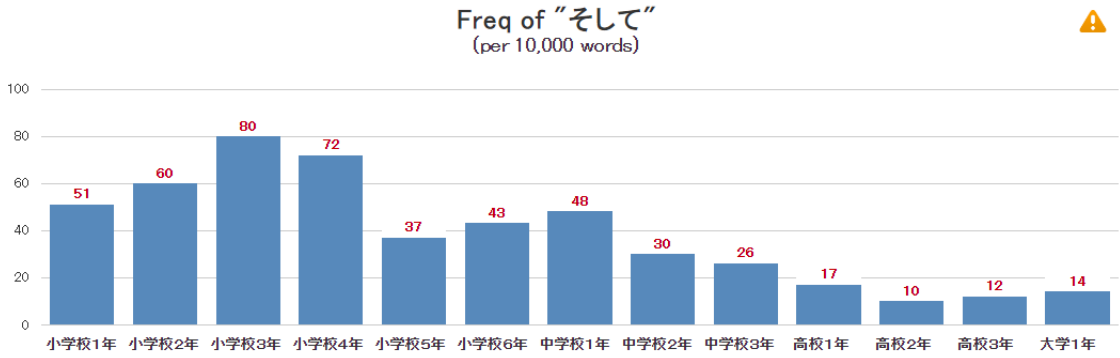


図 13「そして」の学年別頻度変化（1 万語あたり調整頻度）

若干のぶれはあるものの、「そして」の頻度は、およそ小学生で多く、中学生・高校生・大学生になるにつれて減少していく。詳細な分析は本稿の狙いを超えるが、今回の結果に限って言えば、低学年の子どもほど、「そして」を用いて、内容の連続性を明示的に強調しがちであるように思える。砂川（2020）は、L1 児童作文に「接続助詞などで節が連ねられ、だらだらと長く続いて脈絡が読み取りにくい文」が多いことを指摘しているが（4.2 節参照）、小学生による「そして」の多用も砂川の指摘とも呼応する。

３．２．７　特徴語検索

特徴語検索とは、対照群（レファレンス）と比較した場合に、調査対象群（ターゲット）の側で顕著に過剰ないし過少に使用される語を調べる目的でなされる。ターゲット側での過剰と過少は、それぞれ、レファレンス側での過少と過剰を示す。こうした研究では、一部対全体（例：「小 1」対「小 1～大 1」）を比較することが多いが、性質の異なる 2 群に焦点を当て、一部対一部（例：「小 1」対「大 1」）を比較することもある。

ターゲットとレファレンスの指定は、検索条件設定パネルで平易に実行できる。以下は、ターゲットに小学校 1～3 年生を、レファレンスに高校生～大学生を指定し、トピックは両課題、語の単位は書字形とした場合の設定例である（図 14）。

図 14 特徴語検索の検索条件指定画面

以下は、書字形（左側）および語彙素（右側）を単位として、高大生と比較した場合の小学校低学年児童の特徴語を抽出した結果である（図 15）。

Chi2　Log-Likelihood

過剰使用		過小使用	
Word	Statistic	Word	Statistic
いい	121.71	の	81.18
たら	115.33	2人	70.32
そして	98.52	に	49.90
いっ	66.70	しかし	42.67
よん	59.99	階	39.38

過剰使用		過小使用	
Word	Statistic	Word	Statistic
言う	182.30	二人	52.62
そして	122.60	然し	35.83
さん	74.34	二	32.50
御	62.03	階	30.12
ます	52.16	為	27.51

図 15 特徴語検索結果画面の一部（高校生～大学生に対する小学校 1～3 年生の特徴語）

検索結果の上部には、個々の語の特徴度、つまりは、レファレンスに対するターゲット側での過剰・過少度を示す 2 種の統計量の切り替えボタンがあり、選択中の統計量が赤色で表示される。デフォルトでは、カイ二乗統計量（Chi2）の降順で過剰使用語と過少使用語が表示されている。

書字形での分析によると、小学校低学年児童は、前述の順接接続詞「そして」のほか、仮定形助動詞の「たら」や、「言う」の活用形を過剰使用する一方、所有格・目的格を示す格助詞の「の」や「に」、逆接接続詞の「しかし」、また、登場人物を代名詞的に指し示す「2人」などを過少使用することがわかる。また、語彙素での分析によると、低学年児童は、上記の傾向に加えて、丁寧さを示す接辞の「御～」、丁寧形語尾の「ます」、丁寧な人物呼称である「～さん」を過剰使用し、行為の目的を示す「為」などを過少使用している。実際の研究では、こうして得られたキーワードを再度 KWIC 検索や共起語検索にかけ、これらの語の用法を詳しく確認していくことが必要である。なお、統計値については、対数尤度比の使用も可能である。対数尤度比は、対数を使用して頻度の情報を圧縮するため、一部対全体比較のように、ターゲットとレファレンス間でデータ量の差が大きい場合に使用が推奨される。

４　データの概要と研究展望

４．１　L1 日本語発達研究での利用

書き言葉の L1 日本語の発達研究に関しては、語・連語・文法項目などの使用状況の学年別変化の調査が想定される。以下に示すのは、JASWRIC に含まれる 13 学年分の「鍵」作文の中から、2 年間隔で、小 1、小 3、小 5、中 1、中 3、高 2、大 1 の作文サンプルを抽出した結果である（校閲なしの素起こしテキストを表示している）。

(1) ケンは､かぎがないので､ちゃいむをならしましたが､マリは､でてくれませんでた。マリは、ねているので、きこえませんでした。ケンは、きんじょから、はしごをもってきてしまいました。けいさつの、さいれんに、マリは、おきてケンの、ほうをみたら、ケンが、マリのほうをゆびさして、びっくりしました。（小 1、G01_001）

(2) マリはよるの 12 じだからぐっすりねむりについていました。だからはしごでまどからはいろうとしたらけいかんがきてケンをちゅういしてケンがおえりてきてじじょうをはなしたらマリがおきてけいかんがなっとくしておわり。（小 3、G03_001）

(3) 家のチャイムをならしてもだれも出てきません。「2 階でねているのか」と思ってマリをいくら呼んでも聞こえていません。しかたなくケンは庭のはしごを使って家のカーテンが開いている窓をたたこうとしました。「あともう少し」というところで警官にあやしまれ、事情聴取されました。わけを話すと警官も分かってくれたらしく、その声が 2 階にも聞こえたらしくマリが気づいてくれました。（小 5、G05_001）

(4) ケンは言いました。「どうしよう。今日はマリちゃんと今後について話し合う予定なのに…」そこでケンは大声でさけんでマリに気づいてもらおうとしました。しかしぐっすりねてしまっているマリには届きません。そして考えあぐねた結果、自分のカバンからはしごを取り出しました。そうこのやけに小さなカバンはまほうのカバンだったのです。そしてそのはしごをやねにかけ、登ろうとしたその時、「何してるんですか」と声をかけられました。ふりむくとそこには警官が。「あやしいものではないんです、ただ鍵を…」とケンは説明しましたが、警官は、「分かっています。ただ私の話を聞いて下さい。あなたが結婚しようとしている「マリ」こと佐藤魔利は、結婚さぎ師です。佐藤に新たなカモができたと分かったので調査していました、あなたのことです。」「ええっ！？」「あのマリちゃんが？」「はい。今後について話しあうと言って金をまき上げる、彼女のさぎの方法です。」「そうだったのか…ありがとうございます。今日鍵を忘れていたおかげで金を取られずに済みました…ありがとうございました、」「あのマリがさぎ師だったなんて…」二人がほほえみあった時、窓が開き、マリが「何

してるのケンちゃん、おそかったわね。明日のピクニック、用意してるわよ。」「では明後日にもう一度うかがいます」と言い残し、警官は去って行きました。（中 1、G07_001）

(5) マリに声をかけたが、マリはねていたので反応がありませんでした。しかたなく、はしごを使ってあいている二階のまどから中に入ろうとしましたが、警官が勘違いして、呼び止められました。その後警官と和解し、その音でマリも起きてしまいました！（中 3、G09_001）

(6) 家の二階にいるマリに助けを求めますが、寝ているため、返事がありません。そこでケンが、はしごを使い、二階のマリに気づいてもらおうとしたところ、不審に思った警官にとがめられました。ケンが事情を説明し、警官の誤解を解いていると、マリが起きてきて、外のようすを窓からのり出して確かめました。（高 2、G11_001）

(7) 夜中だったのでインターホンを鳴らしても、家の中にいるマリは気づいてくれませんでした。ケンはマリの寝室の窓が空いていることに気付いたので「開けてくれ～！」と呼びかけましたがマリは返事をくれません。なぜならマリはぐっすり眠っているからです。ケンはしばらく呼び続けましたが、庭にはしごをおいてあったことを思い出しました。はしごを開いている窓に立てかけ登っていると、たまたま通りかかった警察に呼び止められました。散々説明しても納得してくれませんでしたが、その話し声でマリが目覚め、ケンは無事、家に入ることができました。（大 1、G13_001）

これらを一瞥するだけで、表記・語彙・文法などの各面において、L1 の子どもの書く日本語の質が連続的に変化していることが確認できるであろう。

また、学年間の変化は、計量的にも議論することができる。図 16 は、2 種のストーリーライティング作文を統合した際の 1 人あたりの総語数と、句読点・普通名詞・動詞・形容（状）詞・副詞・助詞・漢語・外来語頻度を自然対数に変換し、学年別の変化パタンを整理したものである。

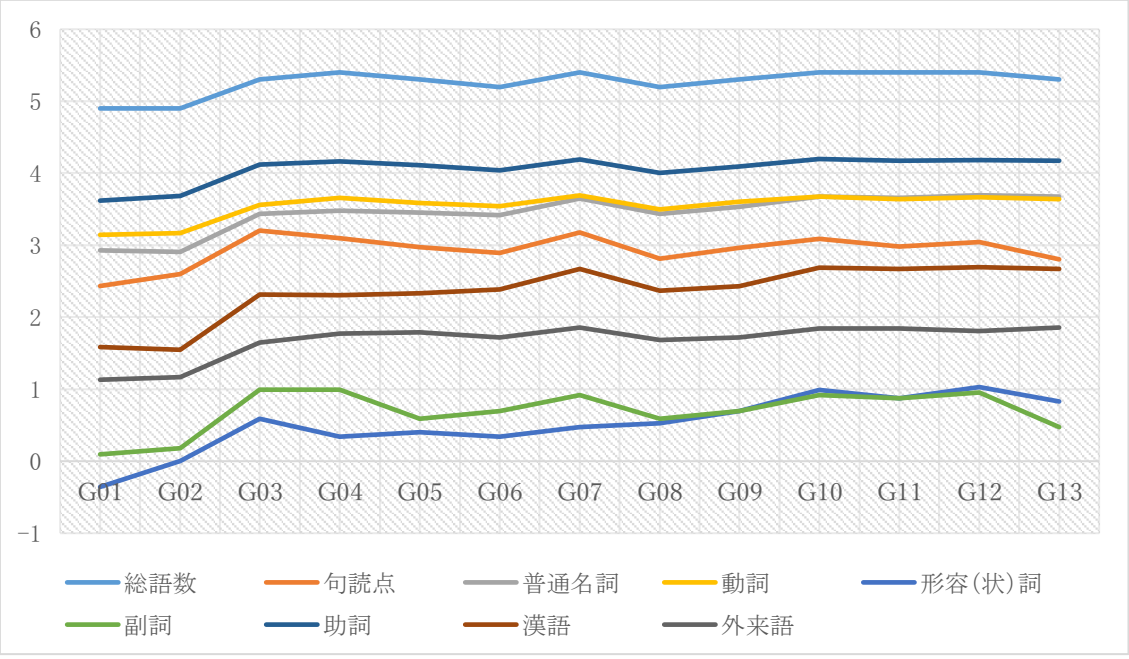


図 16　13 年間の総語数および各種頻度の変化（自然対数変換）

13 学年中、データ量が最も少ない中学校 1 年生 1（G07）の値が高めに出る傾向があるが（G07_001 の「鍵」作文をはじめ、いくつかの作文が例外的に長く、このことが全体の値に影響していると思われる）、その点を除いて考察すると、(1) 小学校 1～2 年生（G01～G02）間では、形容（状）詞・句読点頻度が増加しているが、そのほかの指標における変化の度合いは小さいこと、(2)小学校 2～3 年生（G02～G03）間ではほぼすべての指標で増加が認められること、(3) 小学校 3 年生～大学 1 年生（G03～G13）の間では、形容（状）詞が（相対的に）線型的な増加を見せ、副詞が高校 3 年生～大学 1 年生（G12～G13）間で減少傾向を示す一方、そのほかの指標は大きく変化しないこと、などが確認できる。これらは、L1 日本語の書き言葉の各面の発達が単純な線型的伸長としてとらえにくいことを示しているだろう。また、小 2 と小 3 の間に見られる相対的に大きな伸長は、書き言葉の L1 発達における「発達の節目」がおよそこの時期にある可能性を示唆しているようにも思える。

同じデータを使えば、品詞や具体的語彙項目の頻度変化に限らず、誤用や不適切な用例の出現度の変化を探ることも可能である。従来、L1 日本語の作文研究では、作文技術の巧拙を議論することはあっても、第二言語習得論（SLA）的な意味合いにおいて、日本語としての正用性や誤用性を議論することは少なかった。この点に関して、砂川（2020）は、「児童・生徒作文コーパス」に含まれる小学校 2 年生および 3 年生の作文を分析し、文法や語彙に関するはっきりした誤用は少ないものの、繰り返しや、「だらだら文」が多く見られることを明らかにしている。また、表現上、問題となりうる事例をタイプ別に分類したところ、文法（25.9% : 助詞・ヴォイス・従属節の誤用など）、表現（21.3% : 繰り返し、接続関係・指示対象の不明瞭性、こなれない表現など）、話し言葉使用（19.3% : 接続詞、各種語彙など）、だらだら文（12.2% : 接続助詞による複数の節の連続）、ねじれ文（10.7% : 主述照応の不備）などが多かったと述べている。また、砂川（2022）では、同コーパスに含まれる小学校 2 年生、小学校 5 年生、中学校 2 年生の 3 群の作文を分析し、「不具合」な箇所の合計数が 40→25→12 のように変化することが示されている。砂川の研究で採用された分析の手法は、そのまま、JASWRIC のデータにも適用が可能である。

４．２　L2 日本語習得研究での利用

共通のプロンプトと条件で産出を集めていることから、JASWRIC のデータは、I-JAS のデータと組み合わせて統合分析にかけることもできる。I-JAS にも、年齢分布に配慮した 50 人の成人母語話者（20 代：19 人、30 代：14 人、40 代：13 人、50 代：4 人）のデータが用意されているが、JASWRIC を使うことで、より多様な母語話者データを参照基準として学習者の日本語の諸相を探ることができる。とくに、低学年の母語話者児童と初級学習者、高年次の母語話者と上級学習者の比較は多くの研究上の可能性を有する。

以下、低学年児童（小学校 1 年生）と初級学習者（トルコ語母語、J-CAT スコア=101）、また、高年次生（大学 1 年生）と上級学習者（中国語母語、J-CAT スコア=320）の作文を順に比較してみよう。

(8) それでケンとマリがちずをみていたら、そのあいだに、いぬが、ばすけっとのなかのさんどいっちを、ぜんぶ、たべてしまいました。それで、ケンとマリがぴくにっくにいって、ついたらばすけっとのなかをみてみたら、ばすけっとのなかからいぬがでてきました。それでばすけっとのなかをみたらなにもありませんでした。（G01_011）

(9) マリさんとケンさん食べ物をピクニックに行くために作ります。地図を目るときいぬはバスケットの中にあるたべものをたべましたマリさんとけんさんはわからない時行きました。ピクニックをする所へいってバスケットのなかに見ました。たくさんびっくりしました。（TTR16）

小学校 1 年生の作文は、大半が平仮名で書かれているため、非常に幼い印象を与えるが、

言語的な問題はほぼ皆無で、継続アスペクト（みていた／している）、補助動詞用法（みてみたら／してみる／してしまう）、複合動詞（でてきました）など、日本語教育で難度が高いとされる表現も自然に出現している。一方、初級学習者の動詞使用は、「目る（→見る）」や「見ました」のように単純時制に偏重しており、細かな文法・表現のエラーも数多く認められる。これらは、母語話者の初期段階と学習者の初級段階が一定の重なりを持ちつつも、質的な異なりを有することを示唆するだろう。

次に、高年次生と上級学習者を比較する。

(10) 彼らはサンドイッチやリンゴをバスケットに入れ、ピクニックの準備をしました。しかし、彼らが地図を確認している隙に犬がバスケットの中に忍び込んでしまいました。何も知らないケンとマリはそのバスケットを持ってピクニックに出かけます。着いた先でバスケットを開けた途端、中から犬が飛び出し、二人はとても驚きました。バスケットの中身は、犬に食べられてしまっていて、二人は残念に思いました。（G13_004）

(11) ずっと前から計画してきたから、二人とも楽しみにしています。二人は犬を飼っていますが、犬を連れていくわけにはいきません。実は二人がどこかに行けばいいかと迷っていますので、ちょっと地図を見て場所を決めているうちに、犬はこっそりとバスケットの中には入りました。それでは、二人は歌を歌ったり、笑ったりして、手を握って一緒にピクニックに行きました。昼になると、色々遊んだ二人ともお腹が空いて、バスケットを開けて昼ご飯を食べようとしていたとたん、犬はバスケットの中からどっさりと飛び出して、二人ともびっくりしました。すると、バスケットの中を見ると、用意してきたおいしいサンドイッチもリンゴも食べられてしましました。ああ、本当についていないですね。(CCM35)

大学生作文も動詞の形態的多様性によって特徴づけられるが（継続相：確認している、受け身：食べられて、複合動詞：忍び込んで・飛び出し、テシマウ形：忍び込んでしまいました／食べられてしまっていて）、上級学習者もこの点で大きな違いはなく、多様な動詞形を使用している（継続相：しています・飼っています・決めている・食べようとしていた、受け身：食べられて、複合動詞：連れていく・飛び出して、テシマウ形：食べられてしまし（い）、テクル形：計画してきた・用意してきた）。一方、いくつかの違いも認められる。ここでは、上級学習者作文の問題点を 3 点指摘しておきたい。1 点目は接続表現の過剰使用である。学習者はストーリーの流れを強調しようとして「それでは」「すると」「昼になると」「見ると」といった接続表現を多用するが、大学生作文に同様のつなぎ言葉は出てこない。学習者が使用した接続表現の多くは本質的に不要なもので、削除が妥当である。2 点目はコロケーション上の不整合である。学習者は「どっさりと」「飛び出（す）」と書いているが、2 つの語は通例つながらない（「とっさに」との混同か？）。3 点目は「途端」の不適切な使用である。「途端」は、「何かをする・何かが起こると、ほとんど同時に、急激な変化・因果関係のある事柄・偶然一致の事柄などが起こることを表し、習慣性のない過去の出来事の客観的な描写・事情説明をする」もので（『研究社類義語使い分け辞典』）、前件には時間の特定された具体的・特定的なイベントが来る。この点に関して、母語話者用例にある「バスケットを開けた」＋「途端」という組み合わせは自然だが、学習者用例にある「食べようとしていた」＋「途端」の組み合わせは不整合である。

このように、JASWRIC に含まれる幅広い母語話者データを併用することで、I-JAS を用いた学習者研究・習得研究はさらなる広がりを見せるだろう。また、I-JAS に含まれる 1,000 名の学習者作文と、JASWRIC に含まれる 700 名の L1 作文があれば、大規模な計量的比較研究も可能になると思われる。この方向性については稿を改めて詳しく論じることとしたい（石川、2022）。

５．まとめ

以上、本稿では、JASWRIC の構築過程と、集められたデータの概要について報告を行った。JASWRIC は公開型のコーパスであるため、分析結果を第三者が検証することができる。また、データの使用者は、著作権法の定める範囲において、それぞれの研究目的に沿ってデータを加工・修正したり、新たな情報を付与したりすることもできる。さらには、同じプロンプトを使用して独自のデータ収集を行うことで、リサーチコミュニティ全体で、相互比較可能なコーパスデータのさらなる質的・量的拡大を図っていくことも可能であろう。

コーパス言語学の普及が進む中で、これからのコーパスは、従来のような閉じた完成物ではなく、外部に開かれ、他者を巻き込み、逐次的に変化・発展していく動態的資料に変容していくのではないかと予想される。JASWRIC が、こうした新しいタイプのコーパスとして、多様な研究者・研究機関とのコラボレーションを触発し、L1/L2 の日本語研究の進展に貢献することを期待したい。

注

(1) 本研究は、第一筆者が研究企画・データ収集統括・大学生データ収集・収集データ整理・検索システム開発・データ分析を担当し、第二筆者以下が小学生・中学生・高校生のデータ収集の実務を担当した。

(2) 本稿で示した分析実例は、2022 年 8 月 1 日現在の JASWRIC のベースデータ（V1.0）に基づく。今後、形態素解析データの修正や JASWRIC *Online* の検索仕様の修正などが行われた場合、結果が変動する可能性があるので留意されたい。

謝辞

本コーパスの開発は、子どもの作文収集を試みた先行研究に多くを負っている。坂本真樹氏・冨士原紀絵氏・宮城信氏の 3 氏には、各氏の構築されたコーパスの内容に関する照会に対して、丁寧なご教示をいただいた。今田水穂氏からは、冨士原氏・宮城氏らと構築されたコーパスの最新の語数について情報をいただいた。砂川有里子氏からは、「児童・生徒作文コーパス」のデータを用いて開発中の「問題例検索システム（仮称）」についての情報をいただいた。各氏に深く御礼を申し上げる。また、本コーパスの最大の特徴は、I-JAS のストーリーライティングタスクのデザインを踏襲してデータ収集を行った点にあるが、プロンプトの使用を快諾くださった I-JAS 開発者である迫田久美子氏に改めて深く感謝したい。本コーパス開発を着想する直接のきっかけになったのは、I-JAS のプロンプトを用いて国内児童の産出を収集・分析された松隈杏梨氏の研究（2021）であった。類似したコンセプトに基づくデータの収集についてご理解をくださった松隈氏、ならびに松隈氏の指導教員である丸山岳彦氏に御礼申し上げる。

本研究は、パフォーマンス評価データの体系的収集をテーマとする第一筆者の科学研究費（20H01282）プロジェクトの成果の一部である。また、神戸大学附属学校部校種間連携部門プロジェクト「幼児期から児童期における発達・教育研究」と連携して実施されたものである。関係各位の支援に感謝申し上げる。

文　献

阿部藤子・今田水穂・宗我部義則・冨士原紀絵・松崎史周・宮城信（2017）「児童生徒の『手』作文に於ける経年変化の計量的分析：1992 年と 2016 年の作文を比較して」『言語資源活用ワークショップ発表論文集』*1*, 234–247. http://doi.org/10.15084/00001478

石川慎一郎（2021）『ベーシックコーパス言語学』（第 2 版）. ひつじ書房.

石川慎一郎（2022）「L2 日本語学習者を対象とした中間言語対照分析における参照基準の拡張―『多言語母語の日本語学習者横断コーパス』（I-JAS）と『小中高大生による日本語絵

描写ストーリーライティングコーパス』（JASWRIC）の連動分析の試み―」『計量国語学会第 66 回大会予稿集』43-47.
今田水穂（2020）「『児童・生徒作文コーパス』形態論・係り受け情報データ Ver.1.6」.
今田水穂・宮城信（2020）「学校課題作文コーパスの構築」『言語資源活用ワークショップ発表論文集』*5*, 103-113. http://doi.org/10.15084/0000314
小磯花絵・居關友里子・柏野和佳子・角田ゆかり・田中弥生・宮城信（2020）「子どもの会話コーパスの構築に向けて」『言語資源活用ワークショップ発表論文集』*5*, 157–163. http://doi.org/10.15084/00003155
国立国語研究所（1989）『児童の作文使用語彙（国立国語研究所報告 98）』東京書籍.
坂本真樹（2010）「小学生の作文コーパスの収集とその応用の可能性」『自然言語処理』*17*(5), 75–98.
迫田久美子・石川慎一郎・李在鎬（2020）（編）『日本語学習者コーパス I-JAS 入門: 研究・教育にどう使うか』くろしお出版.
島村直己（1987）「児童の漢字使用：課題作文の漢字含有率から」『研究報告集 8 (国立国語研究所報告 90)』77–94. http://doi.org/10.15084/00001105
鈴木一史・棚橋尚子・河内昭宏（2011）「作文コーパスからみる生徒の使用語彙」『特定領域研究「日本語コーパス」平成 22 年度公開ワークショップ（研究成果報告会）予稿集』343–350.
砂川有里子（2020）「『児童・生徒作文コーパス』に見られる問題例の調査報告―低学年児童の場合―」『日本語教育連絡会議論文集』*33*, 127–135.
砂川有里子（2022）「『児童・生徒作文コーパス』を用いた長文の文構造調査」『日本語習熟論学会第 1 回大会研究発表予稿集』17–24.
田中美也子（1998）「作文能力発達に関する縦断的研究その三：段落意識形成の過程に関する一考察」『語学文学』*36*, A1–9.
永田亮・河合綾子・須田幸次・掛川淳一・森広浩一郎（2010）「作文履歴をトレース可能な子供コーパスの構築」『自然言語処理』*17*(2), 51–65.
成田信子・宗我部義則・田中美也子（1995）「作文能力発達に関する縦断的研究その一：小学生から大学生に至る同題作文の分析」『国語科教育』*42*, 183–192.
冨士原紀絵・宮城信・松崎史周（2016）「児童生徒作文の基礎的研究：児童生徒作文コーパスの構築と活用」『子ども学研究紀要』*4*, 9–20.
松隈杏梨（2021）「JSL 児童生徒の接続詞使用における課題 ―I-JAS 学習環境別データの分析と調査の実施から―」学習者コーパス研究会（2021 年 12 月 26 日実施）発表資料.
宮城信・今田水穂（2018）「『児童・生徒作文コーパス』を用いた漢字使用能力の発達過程の分析」『計量国語学』*31*(5), 352–369.
村上博之・田中美也子（1997）「作文能力発達に関する縦断的研究その二：同題作文における漢語表現の発達」『国語科教育』*44*, 105–114.

参考 URL

「小中高大生による日本語絵描写ストーリーライティングコーパス」（JASWRIC）
https://language.sakura.ne.jp/jaswric/