村田裕美子

ドイツ語話者の書き言葉コーパスの開発

はじめに

ドイツ語圏の日本語教育学研究においてデータに基づいた定量的研究が少ない背景には、ドイツ語を母語とする日本語学習者のコーパスがこれまで存在しなかったことが影響していると考えている。こうした課題を解決すべく、筆者は、2014 年からドイツ語母語話者のコーパス開発を進め、第一弾として「ドイツ語話者話し言葉コーパス（以下、GLJ コーパス）」をウェブシステムを通じて公開させた[1]（村田/李 2017）。現在は書き言葉コーパスの開発に取り組み、GLJ コーパスと同様に公開を目指している。

コーパス開発の目的は、(1)ドイツ語圏の日本語学習者の言語使用の実態を明らかにすること、さらに、(2)その研究成果を日本語教育の現場に還元することである。また、複数のコーパスを開発する目的として(3)多様な言語研究のニーズに少しでも応じたいと考えているからである。本稿では、2020 年現在開発中の書き言葉コーパスについて、そのグランドデザインを示すと同時に、どのような調査研究ができるか、開発中の書き言葉コーパスを用いたケース・スタディを示す。[2]

研究背景と意義

学習者コーパスの進歩

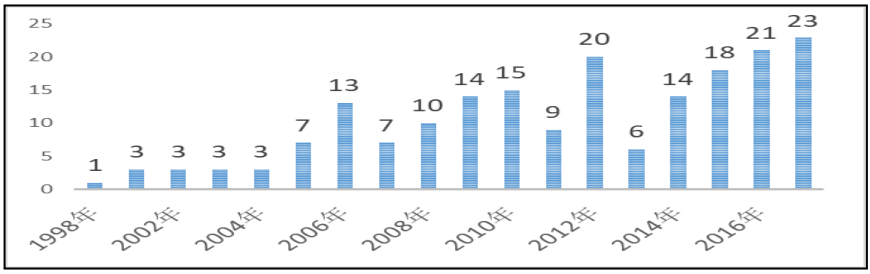
学習者コーパスとは、外国語学習者が産出した言語のコーパス、電子テキストデータベースのことであり（Granger 1998: xi）、第二言語学習者、あるいは外国語学習者が学習言語で実際に話したり、書いたりしたものをテキストとして集めたものである。1990 年代、PC などの技術的な発展により、個人でもコーパスが容易に扱える環境が整ったことから、英語教育の分野でコーパス開発が進み、現在は、日本語教育のほか、第二言語習得や外国語教育の分野で、コーパス開発とコーパスを用いた研究が広まり、言語教育の研究や現場への還元に貢献している。学術情報検索のデータベース（CiNii）で「日本語」「学習者」「コーパス」を検索語に入れて検索すると、2018 年 11 月までに 202 件の論文が該当した。一番古い論文は 1998

[1] ドイツ語母語話者話し言葉コーパスの詳細はコーパス情報源にある URL を参照のこと。

[2] 本稿は 2018 年の Japanologentag において発表した内容に加筆・修正を加えたものである。

年のものであるため、1998 年から 2017 年までの 20 年間でどのぐらい論文数が増えたのかを図 1 に示す。1998 年から 2007 年までは 40 本の論文が、2008 年から 2017 年までは 3 倍以上の 150 本の論文が公開されており、コーパスを用いた研究が広まっていることを示している。加えて、この 202 件の論文のなか、ドイツ語母語話者のデータを扱っている論文は、管見の限り 1 本のみであり、これも多言語データのうちの 1 言語として扱ったものである（cf.奥野/リスダ 2015）。これは、英語、中国語、韓国語を母語とする日本語学習者を扱う論文に比べると圧倒的に少ない。その理由としては、先にも述べたとおり、コーパスデータとして公開されているものに英語、中国語、韓国語以外の学習者データが少ないことが影響しているといえるだろう。

図 1「日本語/学習者/コーパス」に関する研究の 20 年間の広がり



学術情報検索のデータベース（CiNii）の検索結果から

学習者コーパスの役割

Granger（1998: 4）は、第二言語習得研究の目標は、外国語や第二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であり、そのためには学習者の運用データのような実証的データが必要であると述べている。さらに、その実証データとなる学習者コーパスを用いた研究では、たとえば、(1)学習者がどのような場合にどのような誤りをするのか、学習者の苦手な領域を数値で算出し、客観的に特定すること、(2)学習者コーパスを母語別に比較することで、母語の干渉によるものなのかを明らかにすること、(3)母語話者のデータと比較することで、学習者の過剰、あるいは過少使用の実態を調べること、(4)習熟度ごとにデータを比較することで、中間言語の実態を調べることなどが可能となる。

こうした研究から得られた成果は、第二言語習得研究に貢献できるだけでなく、学習者の言語実態を正確に把握できることで、学習者のニーズに対応した教材開発にも役立つと言われている（Granger 1998: 17）。

公開されている学習者コーパス

すでに公開されている主なコーパスについては、村田/李（2017）でまとめているため、ここでは簡単に触れておく。学習者コーパスには、話し言葉コーパスと書き言葉コーパスの2種類がある。

話し言葉コーパスでは、「KY コーパス」、「日本語学習者会話データベース」、「中国語・韓国語母語の日本語学習者縦断発話コーパス（Corpus of Japanese as a Second Language, C-JAS）」、「多言語母語の日本語学習者横断コーパス（International Corpus of Japanese as a Second Language, 以下、I-JAS）」、GLJ コーパスなどが公開されている。特徴として、話し言葉コーパスは、対話を文字化することで作成され、主に OPI（Oral Proficiency Interview）の方法で構築したものが多い。

書き言葉コーパスでは、「日本語学習者作文コーパス」、「対訳 DB」、「日本・韓国・台湾の大学生による日本語意見文データベース」、「日本語教育のためのタスク別書き言葉コーパス」、I-JAS などが公開されている。特徴として、書き言葉コーパスは、作文の形式で収集され、母語での対訳つきで構築されるものが多い。さらに、I-JAS においては、同一調査対象者による話し言葉と書き言葉の様々な課題で集めたデータが収録されている点で優れている。

ドイツ語母語話者を対象としたコーパスについていえば、2020 年 4 月現在は、I-JAS がドイツ語母語話者 50 名分を、GLJ コーパスがドイツのドイツ語母語話者 45 名分を公開しているのみである。

問題の所在

先行する学習者コーパスでは、I-JAS の 50 名と筆者が代表で構築した GLJ コーパスの 45 名を除き、ドイツ語を母語とする日本語学習者のコーパスは 2020 年 4 月現在、存在しない。特に、書き言葉に関しては、I-JAS の 50 名のみとなっている。そのため、ドイツ語圏の日本語学習者の言語使用に関する研究、学習者を対象とした誤用研究や習得研究はまだまだ未開拓の部分が多い。

このような現状をふまえ、本研究ではドイツ語話者の書き言葉コーパスを開発し、話し言葉コーパスである GLJ コーパスとともにウェブシステムを通じて公開することを目指す。コーパスを開発し、研究を支援することで、学習者の言語使用の実態を明らかにし、誤用や習得研究を進め、最終的には教育現場の指導法や教材の開発につなげたいと考えている。

本研究が提案する書き言葉コーパス

概要

本研究が提案する書き言葉コーパスは「ドイツ語話者日本語学習者書き言葉コーパス（Written Corpus of German Learners of Japanese、以下 GLJW コーパス）」という名の作文コーパスである。作文のテーマは、「住みやすい国の条件とその理由」とした。データ収集は、2017 年 5 月から行っており、2020 年 4 月までに 77 名分を集め、現在も収集中である。

調査協力者は主にミュンヘンに在住する日本語学習者であり、教育をドイツで、ドイツ語で受けた学習者である。本コーパスの特徴としては、(1)データ収集時に GLJ コーパスと同様、「SPOT（Simple Performance-Oriented Test；以下 SPOT[3]」を実施し、客観テストによる調査を行っていること、(2)初級から上級までの異なるレベルの学生に協力してもらっていること、(3)全員ではないが、同時に OPI を実施し、話し言葉のデータも含まれている点が挙げられる。これをもとに、習熟度ごとの言語実態や同一調査対象者による話し言葉と書き言葉のジャンルの違いから現れる言語実態などの比較検証を可能にしたいと考えている。

収集手順

データの収集手順について説明する。収集手順は、(1)調査の趣旨説明と調査同意書の確認、および署名、(2)課題提示、(3)SPOT 実施、(4)背景調査と作文のフィードバックである。(2)の課題提示の内容では、(a)課題提示はドイツ語で[4]、(b)分量は 400 字から 1000 字まで、(c)提出はコンピューターを使って自宅で書いたテキストをオンラインで、(d)期間は課題提示後 1 週間以内とした。また、(e)リソースとして、辞書やインターネットなどの使用は認めたが、日本人の友人や知人には手伝ってもらわないことを条件とした。(4)の背景調査と作文のフィードバックでは、コーパス公開時に提供する様々な情報をインタビュー形式で収集した。まず、学習者の母語、教育言語、日本語学習歴、日本滞在歴などを聞き、また、作文のフィードバックの際には、学習者が書いた作文を見ながら、難しかったところや辞書を使って調べた語彙や表現などの聞きとりを実施した。

---

[3] SPOT（Simple Performance-Oriented Test）は日本語の自然な発話速度の読み上げ文を聞きながら、解答用紙に書かれた同文の各 1 文につき 1 箇所のひらなが 1 文字分の空欄に穴埋めディクテーションするというテストである（小林 2015）。

[4] Voraussetzungen (Bedingungen) für ein Land, in dem es sich gut/leicht/schön leben lässt (mit Begründung).

コーパスサイズ

ここでは、収集時期、2017 年 5 月から 2020 年 4 月までに収集した 77 名分のデータのサイズを紹介する。

まず、学習者の作文全 77 編を SPOT の得点によって、初級・中級・上級の習熟度別に分類した。なお、本調査での SPOT 得点の解釈は初級 36-55 点、中級 56-75 点、上級 76 点以上として分類した。以下、表 1 でデータの詳細を示す。

表 1　データの情報とサイズ

習熟度	協力者	SPOT 平均	学習時間 平均（ヶ月）	延べ形態素数（平均値）	総文数（平均値）
初級	21	50.43	9.52	5450 (259.52)	426 (20.29)
中級	49	69.59	35.86	17215 (351.33)	910 (18.57)
上級	7	82	41.86	2638 (376.85)	128 (18.28)
合計	77	65.49	29.22	25303	1464

形態素数、総文数、総文字数は、解析辞書の UniDic と形態素解析エンジンの MeCab の解析結果に基づいて算出。延べ語数は記号、空白、未知語は除いて算出。

表 1 では、中央から左に調査協力者情報として、習熟度と人数、SPOT の平均得点、日本語の学習時間を示し、中央から右には作文のサイズとして、延べ形態素数、総文数を示す。表 1 のとおり、現段階では、収集データに偏りがあり、初級データが 11 名分、中級データが 49 名分、上級データが 7 名分となっている。データのサイズをみると、延べ語数の平均は、習熟度があがるにつれて増えていること（一人当たり初級 259.52、中級 351.33、上級 376.85）、ただし、総文数の平均値をみると、中級と上級では文の数がほぼ同じであることがわかる（一人当たり初級 20.29、中級 18.57、上級 18.28）。

サンプル

ここでは、各レベルの学習者が「住みやすい国の条件とその理由」というテーマで書いた作文をサンプルとして一部紹介する。実際にどのような作文がコーパスとして収録されるかがイメージできるだろう。

初級学習者の作文サンプル

国のいい生活の条件がたくさんあります．政治は親切です．経済はよいで、人はお金があります．冬に天気はさむくて、夏に暑いです．町はきれいです．町に公園や図書館や大学や美術館などが色々あります．町の座賃は安いです．国でお祭りがあります．人

は多くないですが、生産的です．人は月曜日から金曜日まで働いて、週末遊んでします．人の中でだれも悪いで、人々は親切です．外国と国の繋がりは平和的です．（後略）

初級レベルの作文の特徴は、存在文「～があります」や名詞文「～です」など単純な文が多い。指示詞などもほとんど使われておらず、文の羅列が特徴である。加えて、キーボードで打つ際の誤りが確認できた。ドイツ語と日本語では、キーボード上の文字の配列が異なることがあり、特に記号を探すのは難しい。上述のサンプルでも句点がピリオドになっていたり、「家賃」が「座賃」になっていたりする誤りが確認できた[5]。

中級学習者の作文サンプル

人生が良い国の条件
それは独立していて、強い経済を持っているべき国です。経済が強くなると、賃金も上昇したり、下層の生活水準も中産階級とだんだん横並びになるので、ほとんどの人々が消費財や、高級品や、ろくなアパートを買えります。それや便利なインフラや選択の自由や言論の自由や意思の自由などは満足している社会にとって必要です。
政府に嫌なことをさせられてはいけないので、国民も政府も人権で法律を守るべきで戦争に参加しないほうがいいだと思います。（後略）

中級レベルの作文の特徴は、文法的な誤りはあるものの、名詞修飾や「ので」、「べき」、「にとって」、「させられる」などの機能語が現れ、文がやや複雑になっていることが確認できる。また、「思います」を使って意見が述べられるようになっていることも初級と区別される特徴といえるだろう。さらに、一例ではあるが、初級であげた作文サンプルに比べると、一文が長くなっているのが確認できる。

上級学習者の作文サンプル

住みやすい国の条件。住みやすい国というのは一体どんなところでしょうか。私自身が住みたいと思うような国は次の条件を満たさなければなりません。
何よりもまず第一には民主的な人権が大切だと強く思っています。なぜかというと私が生まれ育ったドイツの場合は昔凶悪な独裁があって、若いころから民主主義のありがたさがわかっているからです。言論や報道の自由などの基本的な民主主義の原理がなかったら、住みやすい国だと言えません。また、選挙の詐欺がないということも大事だと感じています。ジョージ・オーウェルが著した『1984 年』という有名な小説に出る独裁国家のような国に住むことは私にとって本当に恐ろしい夢のようです。逆に言えば、国民の代表として怖くない政府がある国に暮らしたいです。（後略）

上級レベルの作文の特徴は、書き出しの一文にも工夫がみられるようになっているところである。サンプルでは「どんなところでしょうか。」と読み手に問いかける文で書き始めている。そのほか、文と文、段落と段落

[5] 「家賃」が「座賃」になってしまった原因は、ドイツと日本のキーボードでは、y と z の位置が逆であるからである（yachin と zachin）。このほかにも括弧「」がどこにあるのかわからず、うまく記号を挿入できない例が確認できた。

をつなげるための「第一に」「また」「逆に言えば」などが現れ、さらに複雑な文構造、そして、段落を意識した文になっていることが確認できる。中級で「思います」というかたちで意見を述べる文が出てきていたが、上級では、「思っています」や「感じています」といった表現が使われるようになり、出だしの一文にある「でしょうか」を含め、文末の表現が多様になっているように思われる。

ケーススタディ：ドイツ語話者の書き言葉の言語的特徴

本節では、筆者が開発し、本研究で紹介している GLJW コーパスのケーススタディとして、2017 年 5 月から 2018 年 8 月までの第 1 期に収集した 50 名分の作文にみられるドイツ語話者の書き言葉の言語的特徴について述べる。村田/李（2017）では、GLJ コーパスを用いて、ドイツ語母語話者の話し言葉には、習熟度によって、どのような言語的特徴が現れるのかを明らかにしている。前節であげたサンプルからは、習熟度の差が文の長さや文頭、文末の表現に現れているような印象を受けた。本節では、これまでの結果とサンプルを見たときに感じられた印象をふまえ、書き言葉の場合には、習熟度によって、どのような特徴が現れるのかを統計的な手法を用いて検証する[6]。具体的には、調査 1 では、データ全体に焦点をあて平均文長の比較と語の多様性の比較を行い、調査 2 では、ある言語形式に焦点をあて文頭と文末の表現の違いについての調査を行った。

調査 1：平均文長と語の多様性

背景と意義

調査 1 では、習熟度ごとの平均文長の比較と語の多様性の比較を行う。前節で紹介したサンプルをみると、初級から中級にかけて、一文が長くなっている様子が視覚的に確認できた。村田/李（2017: 9）でも、ドイツ語を母語とする日本語学習者の発話量を習熟度ごとに比較した結果、初級の場合は、1 発話の長さは平均 16.28 語で構成されているのに対して中級では 22.54 語、上級では 30.26 語で構成されていることが明らかになり、レベルがあがるにつれて、1 発話の長さが長くなることを指摘している。一方で、発話の数自体は上級にあがるにつれて少なくなることから、複文や連体修飾のような複雑な構文を使うことで、1 発話が長くなっていることが考えられた。また、村田/李（2017: 9-10）では語種に関する調査も行い、漢語については初級から上級へとほぼ一定の頻度

[6] 村田/李（2017）では習熟度（初級・中級・上級）は OPI のレベル判定に従っていたが、今回は 50 名のなかに OPI を受けていない者も含まれているため、SPOT によるレベル判定を用いた。

で増えていることが、混種語については、上級レベルになってから産出することが確認できた。これらの結果をふまえると、産出に時間をかけることができる作文の場合では、どのような特徴が現れるかを本調査で明らかにする必要がある。

調査方法

調査を行うにあたり、まず、テキストデータの数値化と出現頻度の計算を行った。テキストデータの数値化では、全作文を、UniDic と形態素解析エンジンの MeCab を用いて、形態素解析し、短単位に区切った。学習者の作文には、誤用や誤植もあるが、今回は、修正せずにそのまま扱っている。出現頻度の計算では、各作文の見出し語（レンマ化した語彙素）から延べ語数、異なり語数、総文数、語種の頻度、品詞別出現頻度を計算した。分析方法には、データ数が少ないため、ノンパラメトリック検定の Mann-Whitney の U 検定を用い、Bonferroni の補正により p=.05/3≒0.017, p=.01/3≒0.003 を有意として多重比較を行った。分析には「IBM SPSS 25」を用いている。さらに、データ数が少ないため、水本/竹内（2008）を参考に、効果量 r を加えた。r で示す効果量は、サンプルサイズによって変化しない、標準化された指標である実質的な差を示している。Mann-Whitney の U 検定では効果大：r =.50、中：r =.30、小：r =.10 が基準である。

結果と考察 1：文の長さ（平均文長）

平均文長に関する記述統計量を表 2 に示す。

表 2　平均文長の記述統計量

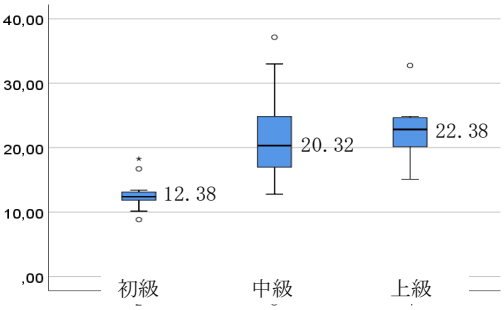
習熟度	中央値	平均値	標準偏差	最小値	最大値
初級（11）	12.38	12.80	2.67	8.82	18.26
中級（32）	20.32	21.31	5.61	12.78	37.14
上級（7）	22.38	22.89	5.5	15.09	32.75
Mann-Whitney の U 検定	初級-中級 0** 初級-上級 0** 中級-上級 1.452				

平均値を用いて、習熟度ごとに Mann-Whitney の U 検定を行ったところ、初級と中級の間（U = 19.00, z = −4.37, p = .0, r = .67）、初級と上級の間では有意であり（U = 2.00, z = −3.31, p = .0, r = .78）、中級と上級の間では有意でなかった（U=92.00, z=−0.73, p=1.45, r=.12）。

次の図 2 は、総形態素数と総文数から割り出した平均文長を習熟度ごとに箱ひげ図[7]で表したものである。数値は中央値である。

作文の一文中に含まれる語数は、初級で 12.38 語、中級で 20.32 語、上級で 23.83 語となり、初級から中級にかけて一文が長くなっていることがわかる。これはサンプルで視覚的に確認できたところであるが、実際の数値でも証明されたといえる。中級と上級の中央値を比較するとそれほど大きな差がないように見受けられるが、中級の箱の大きさが縦に長細いことから、中級は個人差があり、一文の長さが短い者と長い者がいること、上級はそれに比べて、箱が小さくなっており、一文の長さが安定して長いことがわかる。

図 2　習熟度ごとの平均文長



これらの結果から、文の長さは初級から中級にかけて長くなること、中級では文の長さに個人差があること、上級になると個人差が減り、安定して文が長くなるということが明らかになった。このことから、文の長さは、初級から中級にかけての指標であって、中級と上級を区別する指標にはならないということが示された。

結果と考察 2：語の多様性（漢語・異なり語数）

まず、本作文コーパスを用いて、語種に関する調査を行った。その結果、漢語に関しては、習熟度ごとに増加する傾向があり、初級と上級の間で有意な差が確認できた（$U = 8.00$, $z = -2.76$, $p = .012$, $r = .89$）。

[7] 箱ひげ図はデータのばらつき具合を視覚的に示すために用いられる。箱ひげ図では、異なる複数のデータを並べることで、そのデータが散らばっているのか、あるいは集中しているのか、そのばらつき具合を比較することができるのが特徴であり、長方形の箱と上下に伸びる棒で構成されている。長方形の真ん中にある線がデータの「真ん中」を表す「中央値」で、データを順に並べていったときに個数で真ん中に位置する値である。棒の先は最小値と最大値を表す。中央値を中心に、上下にどれくらい「散らばっている」かを確認することができる。箱の大きさと上下に伸びる棒が長ければ長いほど、データが「散らばっている」ということになる。

図 3　習熟度ごとの漢語使用頻度

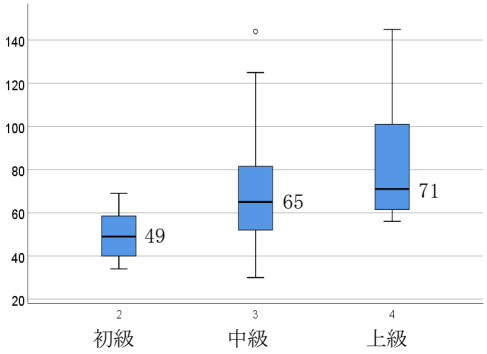


図 3 は、習熟度ごとの平均漢語出現数を箱ひげ図で表したものである。数値は中央値である。中央値からは増加の傾向が確認できるものの、実際は特に中級と上級では、漢語の使用頻度に個人差があり、多く使用する者もいれば、あまり使用しない者もいることがわかった。このような結果をもたらした要因として、作文の課題では辞書の使用を認めていたことが考えられる。

次に、異なり語数に関する記述統計量を表 3 に示す。

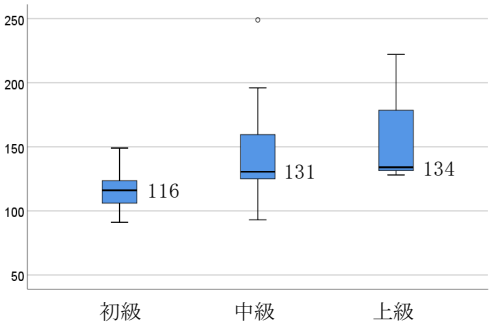
表 3　異なり語数の記述統計量

習熟度	中央値	平均値	標準偏差	最小値	最大値
初級（11）	116.00	117.36	17.36	91	149
中級（32）	130.50	140.19	30.90	93	249
上級（7）	134.00	157.71	38.98	128	222
Mann-Whitney の *U* 検定	初級-中級 .027 初級-上級 .012** 中級-上級 .558				

平均値を用いて、習熟度ごとに Mann-Whitney の *U* 検定を行ったところ、初級と上級の間で有意な差が認められたが（$U = 8.50$, $z = -2.72$, $p = .012$, $r = .64$）、初級と中級の間、中級と上級の間では有意差は認められなかった。

次の図 4 は、習熟度ごとの平均異なり語数を箱ひげ図で表したものである。数値は中央値である。統計的に有意な差は認められなかったものの、箱の大きさに注目すると、レベルがあがるにつれて大きくなっていることから、習熟度の上昇とともに語が多様化する傾向にあることが確認できる。

図 4　習熟度ごとの異なり語数の使用頻度



調査 2：作文の文頭・文末に現れる言語形式の特徴

背景と意義

調査 2 では、作文の文頭と文末に出現する言語形式について、習熟度ごとにどのような特徴があるのかを調べる。

前節で紹介したサンプルをみると、初級では文の羅列が特徴であり、文と文をつなげる接続詞のような機能語は用いられていなかった。しかし、レベルがあがるにつれて、徐々にその使用が目視でも確認できた。また、中級では「思います」が出現していたが、上級では「思っています」「感じています」と表現の種類が増える様子も確認できた。このことから、調査 2 では、接続詞や文末表現に焦点をあて作文の文頭と文末にどのような語が用いられるかを調査することにした。学習者にとって、文頭をどのような語で始め、文末をどのような表現で終わらせたらよいか、適切な言語形式を用いるのは難しい。しかし、文頭の表現には文と文や段落と段落をつなぎあわせる重要な役割があり、文末の表現には書き手の気持ちを表現する重要な役割がある[8]。そのため、作文の指導では、学習者に文頭の接続表現や文末表現の種類を増やし、適切に扱えるように配慮する必要がある。そこで、まずは学習者が書く作文の文頭と文末に、どのような表現が使われているのかを習熟度ごとに明らかにする。

[8] 田中/阿部（2014：17，42，46，175）では、文と文、パラグラフとパラグラフをつなぐ接続表現（「たとえば」「しかし」「したがって」「特に」など）の使用が「good writing」の指標の一つであることを述べている。また、文末のナラティブ表現では、書き手の気持を表す表現であり、使い方によって単調さを避けることができると述べている。

調査方法

Nグラム統計を用いて、文頭と文末にどのような語や表現が出現するかを調べる。Nグラム統計とは、対象となるテキストの中で、連続するN個の文字列の種類、出現頻度を調べるための方法であり、日本語教育でもこれまで利用されている（cf. 山内2004、李/長谷部2017）。

たとえば、「あいうえお。あいう。」というテキストからN＝3個の文字列を検出すると、「あいう」「いうえ」「うえお」「えお。」「お。あ」「。あい」「あいう」「いう。」の8種類の文字列が検出され、そのうち、「あいう」が2回、それ以外は1回の出現であることがわかる。また、文頭（句点のあと）にくる文字列は「。あい」のみ、文末（句点のまえ）にくる文字列は「えお。」「いう。」の2種類である。

本調査では、このNグラム統計を用いて、50編の作文の文頭と文末にどのような語や表現が出現するか、習熟度ごとに検出し、分析する。

結果と考察1：文頭表現

句点のすぐあとにどのような語がくるかを調べるため、句点を含みN = 4で文字列を抽出した結果の上位5を表4で示す。

表4　文頭で出現頻度の高かった上位5（N = 4）

初級	中級	上級
。そして	。たとえ	。また、
。あの国	。そして	。たとえ
。それか	。しかし	。そして
。この国	。それで	。そのた
。とても	。それは	。なぜな

初級は「あの国」「この国」「とても」などが文頭にくることから、接続表現がないまま、文が始まることが多いといえる。これはサンプルで視覚的にも確認できたところである。文をつなげるとしても、「そして」あるいは「それか（ら）」という表現でつないでいくという特徴がある。

しかし、中級になると、指示語「それ」や接続表現「そして」「しかし」が使われるようになる。また、「たとえ（ば）」という語を用いて、具体例が述べられるようになるという特徴も確認できた。さらに、中級と上級では「それで」「そのた（め）」が出現している。「それで」「そのため」は、前件に理由、後件に帰結を述べる表現として、互いに言い換えられることが多い。それぞれの違いとして、「それで」は会話でも使用されるという点、「ため」には「客観的」な性質があり、テキスト的、あるいは文体的な特徴を備えているという点で異なる（庵他2000：217-219、前田2009：148-150）。つまり、「そのため」は「それで」よりも硬い表現、また客観的

な事実を述べるときに使用されやすいということになる。このように、上級にあがるには接続表現のバリエーションを増やしていくこと、「書く」課題に合わせた語を選べるようにする必要があることがわかった。

結果と考察 2：文末表現

句点のすぐまえにどのような語がくるかを調べるため、句点を含みN = 5で文字列を抽出した結果の上位 15 を表 5 で示す。

表 5　文末で出現頻度の高かった上位 15（N = 5）

初級	中級	上級
りません。	思います。	思います。
あります。	あります。	ています。
大切です。	りません。	いと思う。
できます。	できます。	できます。
思います。	大切です。	いいです。
たいです。	いと思う。	でしょう。
いいです。	だと思う。	ないです。
しいです。	ています。	なります。
切ですよ。	ことです。	はずです。
好きです。	必要です。	しょうか。
ができる。	しいです。	い国です。
ですから。	とである。	からです。
何ですか。	ないです。	だと思う。
いですよ。	なります。	べきです。
いります。	たいです。	れている。

初級では、終助詞の「よ」が、初級と中級では「りません。」が多く出現した。ただし、実際のデータを眺めると、初級の「りません。」は「なければなりません。」の多用であり、中級の「りません。」は「なければなりません。」のほかに、「（問題/不可能）じゃありません。」「わかりません。」といった表現としても用いられていることがわかった。さらに、初級と中級では、存在表現・所有表現の「あります。」が多く使われていることが確認できた。実際のデータを眺めてみると、「しごと」、「インフラ」、「権利」、「条件」などと一緒に出現しているが、初級と中級を区別するものはなかった。また、中級から上級にかけては、「思う」を用いて、意見を述べるようになっていることが表から確認できる。さらに、上級では、「ている」表現やモダリティ表現（「でしょう」「はずだ」「べきだ」など）、受身文（「られる」）が用いられるようになり、文が複雑になっていることが明らかになった。サンプルからも、中級で「思います」が、上級で「思っています」や「感じています」などのアスペクト表現が

使われるようになっている様子が見受けられたが、本調査によって統計的にも証明されたといえる。

これらの結果をふまえると、初級では作文のスタイルについて指導する必要があることがわかる。たとえば、終助詞の使用では、「書き言葉」でも誰に宛てて書いたものなのか、目的によっては終助詞が使われることもある。しかし、初級から「話す」課題と「書く」課題の違いや読み手を意識したときの書き方の違いを理解し、適切に使い分けられるようにすることができるとさらに良いだろう。また、中級では文の終わり方、書き手の気持ちを表す表現を増やし、実際に書けるようになっていくことが望まれる。

まとめと今後の課題

本研究では、まず、現在開発中である「ドイツ語話者日本語書き言葉コーパス」を紹介した。本コーパスは、作文コーパスとして開発しており、2020 年 4 月現在のデータ量は、77 名分（初級 21 名、中級 49 名、上級 7 名）である。特徴としては、作文と同時に OPI データ、SPOT 得点も収録しており、すでに公開している「ドイツ語話者日本語話し言葉コーパス」と併せて利用できるようにしている点である。ドイツ語で教育を受けている学習者の書き言葉コーパスとして 70 名分以上を収録しているものは今のところ唯一である。

次に、ケーススタディとして、本コーパスの第 1 次収集期データを用いて 50 編の作文に現れる言語的特徴を習熟度ごとに調べた。1 つ目の調査は、平均文長と語の多様性について習熟度ごとの特徴を統計的な手法を用いて明らかにした。調査の結果、文の長さは、初級から中級にかけて長くなることから、初級と中級を分ける指標であることを示した。また、語の多様性では、漢語の頻度と異なり語数の結果から習熟度があがるにつれて、語のバリエーションが増加する傾向にあること、漢語は個人差によって使用頻度が異なることことが明らかになった。しかし、いずれも初級と上級との間にのみ有意な差が認められた。この要因として、書くときに辞書などのリソースの使用を許可していたことや、時間制限を設けず、時間をかけて産出できたことが影響していたと考えられる。

2 つ目の調査は、文頭と文末の言語形式について、習熟度ごとにどのような特徴が現れるのかをNグラム統計を用いて明らかにした。調査の結果、文頭表現では、中級になって、具体例があげられるようになり、上級になって、客観的でやや硬い表現ができるようになることが確認できた。また、文末表現では、習熟度があがるにつれて、文が複雑になっていく様子が明らかになった。特に上級では、モダリティ表現や受け身表現が使えるようになっていることがわかった。さらに、書き方のスタイルにも違いがみられ、初級では終助詞「よ」が使われることもあった。本調査の結果をふまえ、「話す」課題と「書く」課題の違い、読み手を意識した書き方の指導をする必要がある。

本調査は、サンプルで一例ずつとりあげ、視覚的に比較した際に見受けられた印象を、統計的手法を用いて数値で客観的に証明できたことを示す。コーパス研究では、こうした質から得られた情報を量的に検証したり、あるいは量から得られた情報をもとに質的検証を行ったり、質的研究と量的研究を相補的に用いることが重要である（小林 2010、樋口 2014、李 2017）。そのためにもより良いコーパスの完成が望まれる。

今後の課題として、3 つ示す。

1 つ目は、データサイズを大きくすることである。特に、初級と上級は中級に比べてデータが少ないため、各レベルの収集データが均等になるように集めていく必要がある。

2 つ目は、現在は、単一機関で集めたデータのみであるため、このデータを用いて研究した場合、その結果を一般化できるとは言えない。今後は学習者の国や地域をさらに広げ、多様な学習背景を持つ学習者の作文データを収集していきたいと考えている。

3 つ目は、収集したデータは公開し、言語教育の研究、教材開発など、教育現場に貢献されるべく、一人でも多くの研究者に利用してもらえるよう準備していきたい。そのためにも、まずは筆者自身が、本データを用いて、様々な視点からの分析を行っていく必要があると考えている。

参考文献

庵功雄／松岡弘／中西久実子／山田敏弘／高梨信乃（2000）『初級を教える人のための日本語文法ハンドブック』スリーエーネットワーク.

奥野由紀子／リスダ　ディアンニ（2015）「『話す』課題と『書く』課題に見られる中間言語変異性－ストーリー描写課題における『食べられてしまった』部を対象に－」『国立国語研究所論集』9 号, pp.121-134.

小林典子（2015）「SPOT」李在鎬編『日本語教育のための言語テストガイドブック』くろしお出版, pp.110-126.

小林雄一郎（2010）「テキストマイニングによる学習者作文における談話能力の測定と評価」『STEP BULLETIN』（日本英語検定協会）vol. 22, pp.14-29.

田中真理／阿部新（2014）『Good Writing へのパスポート－読み手と構成を意識した日本語ライティング』くろしお出版.

樋口耕一（2014）『社会調査のための計量テキスト分析－内容分析の継承と発展を目指して』ナカニシヤ出版.

前田直子（2009）『日本語の複文』くろしお出版.

水本篤／竹内理（2008）「研究論文における効果量の報告のために－基礎的概念と注意点－」『英語教育研究』（関西英語教育学会）31, pp. 57-66.

山内博之（2004）「語彙習得研究の方法－茶筌と N グラム統計」『第二言語としての日本語の習得研究』7, pp. 141-161.

李在鎬編（2017）『文章を科学する』ひつじ書房.

李在鎬／長谷部陽一郎（2017）「N-gram を使った文法項目の抽出と学習者コーパスに基づく検証」『計量国語学会機関誌』31（2）, pp. 116-127.

Granger, Sylviane (ed.) (1998): *Learner English on Computer*. Addison Wesley: Longman.

Murata, Yumiko 村田裕美子 u. Lee, Jae-Ho 李在鎬 (2017): Doitsugo washa no hanashikotoba kōpasu no kaihatsu ドイツ語話者の話し言葉コーパスの開発. In: Unkel, Monika (ed.): *Beiträge zum Japanologentag 2015 in München, Sektion Japanisch als Fremdsprache*

(Schriften der Gesellschaft für Japanforschung, Band 2). Köln: Gesellschaft für Japanforschung e. V., pp. 1-19.

コーパス情報源

作文対訳 DB
http://contr-db.ninjal.ac.jp/essay_01.html (2016 年 11 月 15 日閲覧).
多言語母語の日本語学習者横断コーパス (International Corpus of Japanese as a Second Language, I-JAS)
http://lsaj.ninjal.ac.jp/?cat=3 (2018 年 12 月 24 日閲覧).
中国語・韓国語母語の日本語学習者縦断発話コーパス (Corpus of Japanese as a Second Language, C-JAS)
http://c-jas.ninjal.ac.jp (2017 年 11 月 3 日閲覧).
ドイツ語話者話し言葉コーパス (GLJ コーパス)
http://german-opi.jpn.org/ (2018 年 11 月 10 日閲覧).
日本・韓国・台湾の大学生による日本語意見文データベース
http://www.tufs.ac.jp/ts/personal/ijuin/terms.html (2016 年 11 月 15 日閲覧).
日本語学習者会話データベース (横断調査編)
https://nknet.ninjal.ac.jp/nknet/ndata/opi/ (2016 年 11 月 15 日閲覧).

日本語学習者作文コーパス
http://sakubun.jpn.org (2016 年 11 月 15 日閲覧).
日本語教育のためのタスク別書き言葉コーパス
金澤裕之編 (2014)『日本語教育のためのタスク別書き言葉コーパス』ひつじ書房.
KY コーパス
http://www.opi.jp/shiryo/ky_corp.html (2016 年 11 月 15 日閲覧).