

学習者コーパス研究における横断・縦断データ統合の意義： I-JAS と B-JAS をめぐって

石川 慎一郎（神戸大学）[†]

Significance of Integrative Analysis of Cross-sectional and Longitudinal Data in Learner Corpus Research: A Study Based on the I-JAS and B-JAS

Shin'ichiro Ishikawa (Kobe University)

要旨

本論文では、「多言語母語の日本語学習者横断コーパス」(I-JAS) と「北京日本語学習者縦断コーパス」(B-JAS) を用い、共通のストーリーライティング課題 (SW1) における中国語母語の日本語学習者の産出データを分析した。習熟度・学習段階別に 4 つの学習者群に分けて計量的に比較した結果、2 つのコーパスは総単語数および主要品詞頻度の変化において類似したパターンを示した。しかし、各段階を特徴づけるキーワードや、主要品詞・主要動詞頻度に基づく統計的分類においては、異なるパターンが観察された。これらの結果は、今後の学習者コーパス研究において、横断的データと縦断的データを統合的に分析することの意義を示唆している。

Abstract

This paper analyzed the written outputs by Chinese learners of Japanese elicited in the common story writing tasks (SW1) included in the International Corpus of Japanese as a Second Language (I-JAS), which is a cross-sectional dataset, and the Beijing Corpus of Japanese as a Second Language (B-JAS), which is a longitudinal dataset based on the same data collection scheme. A quantitative comparison of the four learner groups, who were classified according to the proficiency levels and the learning stages, revealed that the two corpora showed a similar pattern in terms of the changes in the total number of words and the frequency of major word classes. However, it was also observed that the two corpora presented a different picture in terms of the keywords characterizing each of the learner groups as well as the statistical learner group classification based on the frequency of major word classes and major verbs. These results suggest the significance of integrated analysis of cross-sectional and longitudinal data in future learner corpus research.

1. はじめに

学習者コーパスには、ある一時点において、年齢や習熟度の異なる多様な学習者の L2 産出を一斉に収集する横断コーパス (cross-sectional corpus) と、1 名ないし数名の学習者の L2 学習過程を長期にわたってモニターし、その間、数回 (線形近似や、いわゆる U カーブの検証を行うには一般に 3 回以上が必要とされる : Ployhart and Vandenberg, 2010, p. 97) にわ

[†] iskwshin@gmail.com

たって L2 産出を収集する縦断コーパス (longitudinal corpus) が存在する。

このうち、縦断コーパスは、発達の過程を直接記録し、多くの場合、学習者背景に関する詳細なメタデータが付属することから、発達コーパス (developmental corpus) や濃密コーパス (dense corpus) と呼ばれることもある (Meunier, 2015)。「発達を調査しようとする場合、同じ学習者集団を発達の時系列に沿って追跡していくことが理想」であり (Myles, 2015; 英文文献からの引用は拙訳による。以下同)、縦断コーパスと付属する背景データを組み合わせることで、学習者の行動や周辺環境において起こる様々な変化を加味しながら、個々の学習者の学習過程 (developmental paths) を解明することができる (Vyatkina and Cunningham, 2015)。

一方、縦断コーパスの構築はしばしば困難な取り組みとなりうる。Meunier (2015) も言うように、縦断コーパスの構築は時間がかかり、事前計画を立てにくく、研究アウトプットを出しにくく、短期プロジェクトに比べて助成金を得にくく、さらには、参加者が自然減少 (attrition) する場合も多い。また、縦断と言っても、L2 習得の始点から終点まで、つまり、L2 を習い始めた時点から、母語話者相当の L2 能力を獲得する時点までの全過程を調査することはそもそも不可能であり、多くの場合、「少数の学習者を対象とした、数か月ないし数年程度の L2 発達の記録」とならざるを得ない (Myles, 2015)。

こうした背景をふまえ、縦断コーパスに代わり、横断コーパスを用いて発達を研究する試みも広くなされている。横断コーパスの参加者を習熟度や学年といった観点で群化し、各群 (たとえば、初級・中級・上級など) を比較することで習得プロセスを間接的に観察するのである。こうした手法は、疑似縦断分析法 (pseudo-longitudinal analysis) (Johnson and Johnson, 1999) や、準縦断分析法 (quasi-longitudinal analysis) (Granger, 2002) などと呼ばれる。Myles (2015) は、各群で十分な数の学習者数確保されており、かつ、外的基準による信頼できる習熟度情報が備わっているならば、横断データを用いて学習者の「発達の道筋」 (developmental routes) を跡付けることが可能であると示唆している。

もっとも、ここで留意すべきは、横断データを用いた疑似縦断分析で得られた結果と、縦断データを直接分析して得られた結果の一致が必ずしも検証されていないということである。こうした研究がなされてこなかったのは、相互比較可能な横断・縦断データがほとんど存在していなかったという事情による。実際、学習者の背景や、産出のトピック、また、タスクの内容が異なる横断・縦断データを比較して何らかの異同が明らかになったとしても、データ収集法の差以外の要因による可能性が高く、横断・縦断データの関係性について確定的な結論を導き出すことはできない。ゆえに、横断コーパスと縦断コーパスが学習者の L2 習得研究の資料としてどの程度の同等性を持ちうるのか、あるいは、それぞれの強み・弱みを補うため、どのように両者を統合的に利用すべきか、といった点についての共通理解はいまだ醸成されていないのが現状である。

こうした状況が長く続いてきた中、迫田久美子氏が中心となって開発した「多言語母語の日本語学習者横断コーパス」(I-JAS) (迫田, 2020) をモデルとして、同一タスクで縦断的にデータを集めた「北京日本語学習者縦断コーパス」(B-JAS) がリリースされたことは注目に値する。本研究は、これら 2 種のコーパスから、中国語を母語とする日本語学習者 (Chinese learners of Japanese、以下 CLJ) の L2 習得パターンを計量的に抽出し、その異同を比較することで、横断データと縦断データの関係性について検証を試みたい。

2. 先行研究

日本語学習者コーパスを用いた研究では、研究者の関心対象である文法項目（テンス、アスペクト、授受表現、敬語、など）に焦点を絞り、その習得過程を質的観点から調査したものが多く、一方で、学習者の産出テキストの全体を対象として、L2 習得の総体的過程を量的観点からモデル化しようとした研究もある。ここでは、後者のタイプの研究の例として、筆者の過去の研究の概要を示す。

まず、I-JAS を用いた研究として、Ishikawa (2017) は、中国語・韓国語・英語・トルコ語を母語とする学習者のストーリーライティング産出を調査した。学習者を習熟度別に7段階に区分して比較した結果、習熟度の上昇によって産出語数が増加すること、学習者は助詞・動詞・副詞を過小使用し、名詞・助動詞などを過剰使用すること、初級・中級では母語別の違いが出るが、上級になると母語の差は縮小していくことなどを明らかにした。

石川 (2020) では、対話タスクにおける形容詞の使用状況を分析し、中国語・韓国語・英語を母語とする中級学習者による形容詞の使用量は日本語母語話者の半分程度であること、日本語母語話者の形容詞使用は過去形とくだけた「ちっちゃい」「すごい」などによって特徴づけられることを示した。

石川 (2021a) は、世界の日本語学習者 850 人を J-CAT スコアの 10 点刻みで細かく群化し、ストーリーライティング産出のデータを用いて、習熟度の発達と動詞使用の関係を量的に概観した。その結果、(1)学習者の使用する動詞のトークン数およびタイプ数は母語話者より 22~28%程度少ないこと、また、習熟度別に見た場合、トークン数・タイプ数とも初級から中級にかけて増加し、母語話者を上回る水準に達した後、再び減少する逆 U 字型の分布を見せること、(2)母語話者・学習者・上級者・初級者を特徴づける動詞群が存在すること、(3)学習者の動詞発達過程がおおよそ 4 段階に分かれること、(4)少数の動詞（居る、見る、為る、作る、有る）の頻度で習熟度が予測できる可能性があること、などを示した。また、石川 (2021b) では同様の分析を韓国語母語の学習者データに適用し、石川 (2021a) で観察されたパターンが特定母語の学習者群に対してもほぼ同様にあてはまることを示した。

次に、縦断コーパスを使用した研究として、石川 (2018b) は、迫田久美子氏が中心となって構築した「中国語・韓国語母語の日本語学習者縦断発話コーパス」(C-JAS) を用い、3名の CLJ の発話データを語彙解析した。その結果、(1)トークン数とタイプ数に関しては3名の学習者ともに増加するものの、語彙多様性指標についてははっきりした変化が起こらないこと、(2)上位語は総じて安定的であるものの、初期においては機械的な応答標識が多く、中期・後期になると、次第に文法的に複雑な文を構成する上で必要となる語が多く使用されるようになること、(3)個体の差は発達段階の差よりも強く語彙使用に影響しているが、個体内では発達段階の差も見られること、(4)3名の学習者の語彙習得段階がおおよそ3期に区分でき、それぞれが異なる語群によって特徴づけられることなどを明らかにした。この研究で得られた知見の一つは、縦断データにおいては個人差が非常に大きく、3名のデータの安易な一般化は危険である、ということであった。

そこで、石川 (2018a) では、台湾の東呉大学で構築された縦断コーパスである LARP at SCU を用い、17名の CLJ 大学生が3年半にかけて書いた作文中のタイプ数・トークン数、句読点使用率、品詞使用率、高頻度語彙使用率などを調査した。その結果、初期においては個人差によるばらつきが大きいのが、大学入学後20か月（大学2年生後期）ごろを境界として次第に値が安定するようになり、この時期に日本語の習得が一定の定着段階に達しているのではないかという示唆を得た。

以上の筆者の研究は、いずれでも、横断データか縦断データかのいずれかを用いたもので、両者を同時に使用して結果を比較するということは試みていない。これは、一連の研究で使用した I-JAS、C-JAS、LARP at SCU という3つのコーパスが、それぞれ異なる条件でデータを集めており、相互比較が困難だと判断したためである。

これに対し、特定の文法事項を対象を絞りつつ、横断・縦断データをうまく組み合わせて主張の妥当性を高めている研究も存在する。一つの好例は、趙 (2015) であろう。同研究で

は、縦断型のインタビューコーパスである C-JAS と、横断型のインタビューコーパスである KY Corpus を用い、「テイナイ」の習得過程を検討している。まず、縦断データ分析より、CLJ3 名と韓国語母語の学習者 3 名は、第 1 期から第 8 期にかけて、「テイナイ」の各種用法に関して、ともに、「未完了」「状態」→「属性」→「全面否定」「反復」の順に使用するようになることが示された。次に、横断データ分析により、中国語・韓国語・英語の母語話者は、いずれも、「未完了」「状態」→「属性」「反復」（「全面否定」は高レベルでも少ない）の順に使用するようになることが示された。若干のずれはあるものの、全体として傾向は類似しており、著者は、これらを根拠として、「テイナイ」の用法習得順序に普遍性があることを示唆している。

今後の学習者コーパス研究においては、得られた知見の説得力を高めるため、趙 (2015) のように、横断データと縦断データを併用していくことが望ましいと思われるが、その際、前提として確認すべきは、そもそも、横断・縦断データはどの程度同質であるのか、あるいは、異なるのか、という点であろう。この意味で、高度に統制された横断・縦断データを計量的に比較し、各々のデータの性質を把握しておくことは不可欠であると思われる。

3. リサーチデザイン

3.1 目的と RQ

本研究は、I-JAS と B-JAS に含まれる CLJ のストーリーライティングデータを用い、学習者の段階によって、(1)総語数、(2)主要品詞頻度、(3)特徴語、(4)学習者群の相互関係性、がどのように変化するかを観察することで、横断データと縦断データから得られる結果の一致度を検証し、今後の学習者コーパス研究における横断データと縦断データの統合的使用の指針を得ることを主たる目的とする。本研究で使用するデータは、学習者の母語背景と産出タスクの両方がそろっているため、横断データと縦断データの性質の違いの検証に適したものとなっている。

本研究で具体的に検討しようとする研究設問 (RQ) は以下のとおりである。なお、上記の観点(4)については、主要品詞頻度と、石川 (2021a, 2021b) でも使用した主要動詞頻度を分類の基準とする。

RQ1 総語数の変化に関して、横断・縦断データの同一性はどの程度か？ (総語数)

RQ2 主要品詞頻度の変化に関して、横断・縦断データの同一性はどの程度か？ (品詞頻度)

RQ3 特徴語の抽出に関して、横断・縦断データの同一性はどの程度か？ (特徴語)

RQ4 主要品詞頻度および主要動詞頻度に基づく段階分類に関して、横断・縦断データの同一性はどの程度か？ (段階分類)

3.2 使用するデータ

前述のように、本研究は、I-JAS と B-JAS に含まれるストーリーライティングのデータを分析対象とする。使用したデータは、2023 年 7 月時点における最新版である。両コーパスで実施された対面調査 (インタビュー) は、はじめに、ストーリーテリング (4~5 コマの漫画を見てそのストーリーを口頭で述べる; 約 10 分) を行い、その後、対話 (決められた内容・手順に基づいてインタビュワーが行う質問に答える; 約 30 分)、ロールプレイ (謝罪・依頼という言語機能を含む状況についてロールプレイを行う; 約 10 分)、絵描写 (1 枚絵を描写; 約 5 分) を順次行い、最後に、ストーリーライティング (約 20 分) を行う (迫田・石川・李, 2020, p. 34)。ストーリーライティングでは、冒頭のストーリーテリングで使ったものと同じ漫画イラストを見て、ストーリーをパソコン上で入力する形で実施される (同上, p. 41)。

ストーリーライティング (SW) には、SW1 (ピクニックに関する 5 コマ漫画のストーリーを作る) と、SW2 (鍵に関する 4 コマ漫画のストーリーを作る) の 2 種が存在するが、トピックの影響を可能な限り統制するため、本研究は、SW1 のみを分析対象とする。

3.3 対象学習者

本研究は、学習者の母語背景と学習背景を統一するため、I-JASに参加した1,000名の学習者中、大陸在住のCLJ（CCHおよびCCM）100名と、B-JASのCLJ大学生17名を分析対象とする。なお、両コーパスの参加者は、全員が、J-CAT（聴解・語彙・文法・読解の4観点別の日本語能力測定。各100点、400点満点）と、SPOT（文章穴埋め式の総合的日本語能力測定。90点満点）という2種のテストを受験している。

3.3.1 I-JASのCLJ

I-JASのCLJ100名については、疑似縦断分析を行うため、J-CATの4観点スコアとSPOTスコアの計5種のテストスコアを主成分分析の手法で合成した上で、第1主成分スコアを基準として、習熟度別に4段階に区分する。表1は、5種のスコアの量的概要である。

表1 I-JASのCLJ100名の習熟度テストスコアの概要

変数	人数	平均	不偏分散	標準偏差	最小値	最大値
J-CAT 聴解	100	56.1	227.9	15.1	8.0	86.0
J-CAT 語彙	100	64.8	122.6	11.1	40.0	100.0
J-CAT 文法	100	56.9	232.3	15.2	23.0	91.0
J-CAT 読解	100	51.7	136.5	11.7	11.0	82.0
SPOT	100	73.8	46.5	6.8	57.0	90.0

一般に、世界の日本語学習者の中でCLJの日本語能力は総じて高く、一部は母語話者に準じる日本語力を持っているとされる。ただ、今回のデータに限って言うと、J-CATの語彙セクションを除くと最大値が満点に達しているものはなく、これらのテストは受験者の能力値を正しく測定しているものと推定できる。

5種のスコア内に線形結合している変数はなかった。分析で得られた第1、第2、第3主成分の寄与率はそれぞれ61.3%、13.5%、11.4%であり、第1主成分で分散の6割以上がカバーされていた。また、第1主成分の負荷量の係数はすべてプラスであり（図1）、第1主成分が5種のスコアの代表値になっていることが確認された。

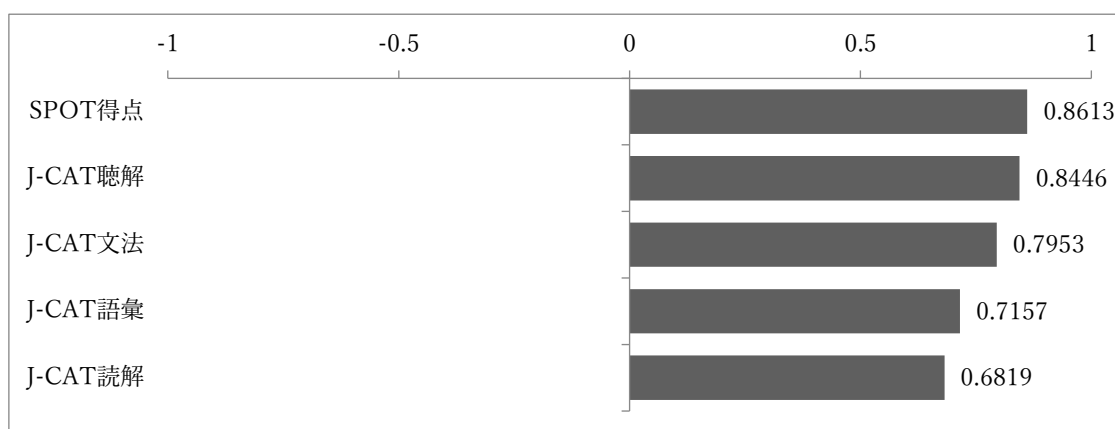


図1 第1主成分の負荷量

個々の学習者ごとに第1主成分得点を求めたところ、主成分得点のレンジは-5.11~+5.29となった。スコアレンジをおよそ均等に分割しつつ、かつ、各群に10名以上の学習者が入るよう調整を行い、表2のような4段階習熟度区分を設定した。

表 2 I-JAS の CLJ100 名の 4 群化

	段階 1 (I1)	段階 2 (I2)	段階 3 (I3)	段階 4 (I4)
主成分得点	-2.0 未満	-2.0 以上	0.0 以上	2.3 以上
学習者人数	12	39	37	12
J-CAT 総合点平均値	168.3	206.2	251.2	299.2
SPOT 平均値	64.0	70.9	77.0	83.2
推定能力水準	中級	中級	中級～上級	上級

J-CAT の総合点に基づいて言うと、段階 1～2 は「日常的な会話をこなすことができる」中級相当（中級～中級後半）、段階 3～4 は「学術的・専門的なコミュニケーションができる」上級の前半に相当する。また、SPOT に基づいて言うと、段階 1～段階 3 は「自然な発話速度で日常的な場面の日本語がある程度理解できる」とされる中級相当、段階 4 は「自然な発話速度で幅広い場面の日本語が理解できる」上級相当ということになる（両テストのスコアの解釈については、李・小林・今井・酒井・迫田、2015 に基づく）。

3.3.2 B-JAS の CLJ

B-JAS の CLJ 大学生 17 名については、データの収集時期を手掛かりとして、4 段階に区分する。B-JAS は、北京の大学で日本語を主専攻とする 17 名の CLJ を対象として、4 年間（2016 年 1 月～2019 年 4 月）にわたって毎年 2 回、計 8 回の縦断的データ収集を行っている。ただし、SW については、I-JAS タスクと独自タスクを交互に課しているため、今回分析対象とする SW1 のデータがとられたのは表 3 に示す 4 回である（いずれもほぼ同じタイミングで 2 種の試験を受験している）。なお、表 3 中の学習時間は、2015 年 9 月に新学期が開始したとして、各時点における通算の日本語学習時間の目安を月数で表示したものである。

表 3 B-JAS の CLJSW1/2 調査回数と学習時間

	段階 1 (B1)	段階 2 (B2)	段階 3 (B3)	段階 4 (B4)
データ収集時期	2016/1	2016/10	2017/9	2018/10
推定学習時間	5 か月	14 か月	25 か月	38 か月
学習者人数	17	17	17	17
J-CAT 総合点平均値	136.2	193.7	224.8	263.1
SPOT 平均値	60.4	71.8	75.2	80.9
推定能力水準	中級	中級	中級	上級

J-CAT の総合点に基づいて言うと、段階 1～3 は中級相当（中級前半～中級後半）、段階 4 は上級の前半に相当し、SPOT に基づいて言うと、段階 1～段階 3 は中級相当、段階 4 は上級相当となる。各段階の平均スコアは、J-CAT では B-JAS のほうがやや低いが、SPOT ではほとんど差がなく、初級・中級・上級という大まかな習熟レベルでとらえれば、I-JAS と B-JAS の 4 段階の能力水準はほぼ同等であると推定される。

3.4 テキストデータの処理

まず、I-JAS の習熟度別 4 群（以下、I1～I4）と、B-JAS の学習時間別の 4 群（以下、B1～B4）の各々について、該当する学習者による SW1 の産出テキストを統合してマージファイルを作成した。その際、各行左端に記載されているコードは除去している。

なお、I-JAS の公開版テキストデータには、丸括弧内に、各種のメモ（「括弧を変更」といった修正記録、「過剰使用」などの学習者の言語使用特性についての注記、「犬の名前」な

どのテキスト理解のための補足情報など)や、学習者の軽微な書き間違いについての修正提案が記載されている。前者については一律で削除し、後者については修正提案を採用した。B-JAS の公開版テキストデータには丸括弧内の追記は見当たらなかったが、I-JAS 側との整合を取るため、同等レベルの誤記については筆者が修正を加えた。また、データ全体に対して、文字種レベルでの若干の統一(人名の「けん」や「まり」はカタカナに、「いぬ」は「犬」に、など)を加えた。

以上の処理後に、I1~4、B1~4 の 8 ファイルを「web 茶まめ」(現代語辞書)で解析し、結果を 1 枚のエクセルシート上に集約した。なお、書字形欄が空データ(補助記号扱いのものを含む)になっている行は削除した。また、語彙素列に含まれる注記(ハイフンで追記された外来語の原語、多義語の用法など)についても一律で削除した。また、RQ3 の特徴語分析のため、語彙素列のみを取り出して、8 種の語彙素単位のテキストファイルを作成した。

3.5 処理手順

RQ1 (総語数)については、I1~4、B1~4 の 8 ファイルから、1 人あたりのトークン数(平均値)を調査し、I-JAS と B-JAS の間で、段階変化に伴う総語数(5 種の補助記号を除く)の変化パターンがどの程度一致しているか確認する。

RQ2 (品詞頻度)については、主要品詞の中から、名詞(普通名詞:サ変可、サ変形状詞可、一般、形状詞可、助動詞可、副詞可)、「部屋」「犬」「サンドイッチ」など)、動詞(一般、非自立可)、「気付く」「飛び込む」「為る」「食べる」など)、格助詞(「の」「を」「に」など)、接続助詞(「て」など)、係助詞(「は」)の 5 品詞に注目し、1 人あたりの使用トークン数(平均値)を調査し、I-JAS と B-JAS の間で、段階変化に伴う主要品詞の頻度変化のパターンがどの程度一致しているか確認する。なお、各品詞の頻度は絶対値に大きな異なりがあるため、ここでは、すべてを自然対数に変換した上で比較を行う。

RQ3 (特徴語)については、I1~4 の各語彙素単位ファイルとそれらの総体ファイル、また、B1~4 の各語彙素単位ファイルとそれらの総体ファイルにおいて、出現する語の頻度を悉皆的に調査して対数尤度比統計量(log-likelihood ratio: LLR)を計算する。今回は、データの絶対量が少なく、通常的手法では十分な数の特徴語が検出できないことから、有意傾向の判定基準となる $\alpha = 10\%$ の下限値となる 2.71 を超えるすべての語を抽出する。上述の理由から、検定反復にともなって必要とされる多重比較補正は行わない。また、頻度の絶対的な差を示す効果量指標として、相互情報量(mutual information)を計算する。特徴語の抽出と、相互情報量の計算には、多言語対応コーパスコンコーダンサである Antconc Version 4.2.0 を使用する。その後、I-JAS と B-JAS の間で、各段階の特徴語が、どの程度、内容的に一致しているかを確認する。

RQ4 (段階分類)では、まず、上述の 5 品詞に、形容詞類(形容詞:一般、非自立可、形状詞:一般、助動詞語幹)、「悲しい」「楽しい」など)、助動詞(「れる」「ます」「た」など)、副詞(「そろそろ」「もう」「がっかり」など)、接続詞(「そして」「しかし」など)、連体詞(「この」「その」など)を加え、合計 10 品詞を第 1 アイテム、学習者 8 群を第 2 アイテムとする頻度表に対して対応分析を実施する。対応分析とは、頻度表の行列の相関を最大化する少数の次元を抽出し、通例、上位の 2 次元をそれぞれ横軸・縦軸とする散布図を描画し、アイテムカテゴリ間の相互関係を可視化する多変量解析手法のことである。ここでは、散布図の質的解釈を通して、I-JAS の 4 段階の位置関係と、B-JAS の 4 段階の位置関係の一致性を確認する。その後、石川(2021a, 2021b)の手法を参考として、頻度上位 30 種の動詞(語彙素)を第 1 アイテム(表 4)、学習者 8 群を第 2 アイテムとする頻度表に対して対応分析を実施し、上記と同様の観察を行う。

表 4 RQ4 で使用する高頻度 30 動詞

高頻度動詞（上位 30 語）
為る、居る、見る、食べる、行く、入る、仕舞う、開ける、飛び出す、作る、来る、持つ、有る、着く、知る、入れる、飛ぶ、出る、成る、思う、開く、分かる、飛び込む、終わる、読む、言う、切る、驚く、つく、遊ぶ

4. 結果と考察

4.1 RQ1 総語数

横断・縦断データである I-JAS と B-JAS において、段階別に、1 人あたり総語数（トークン数）の変化を調査したところ、図 2 の結果を得た。

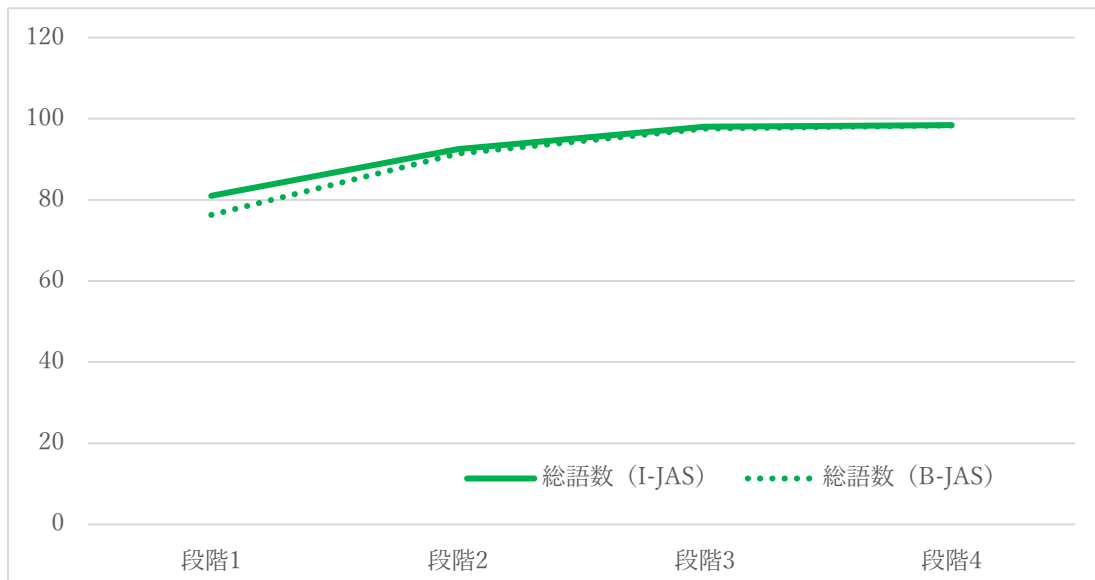


図 2 SW1 総語数 (1 人あたり) の段階別変化

上記より、I-JAS においても B-JAS においても、(1)語数はおよそ 80~100 語の範囲に収まること、(2)段階が進むにつれて総語数が増加すること、(3)語数の増加幅は段階 1→2 の間が最も大きく、以下、段階 2→3、段階 3→4 になるにつれて小さくなること、が確認された。段階 1 において、B-JAS のほうがやや語数が少ないことを除けば、総語数の変化パターンは、I-JAS と B-JAS で高い同一性を示していると結論できる。

4.2 RQ2 品詞頻度

まず、I-JAS と B-JAS において、段階別に、1 人あたりの名詞・動詞頻度（トークン数を自然対数に変換）の変化を調査したところ、図 3 の結果を得た。

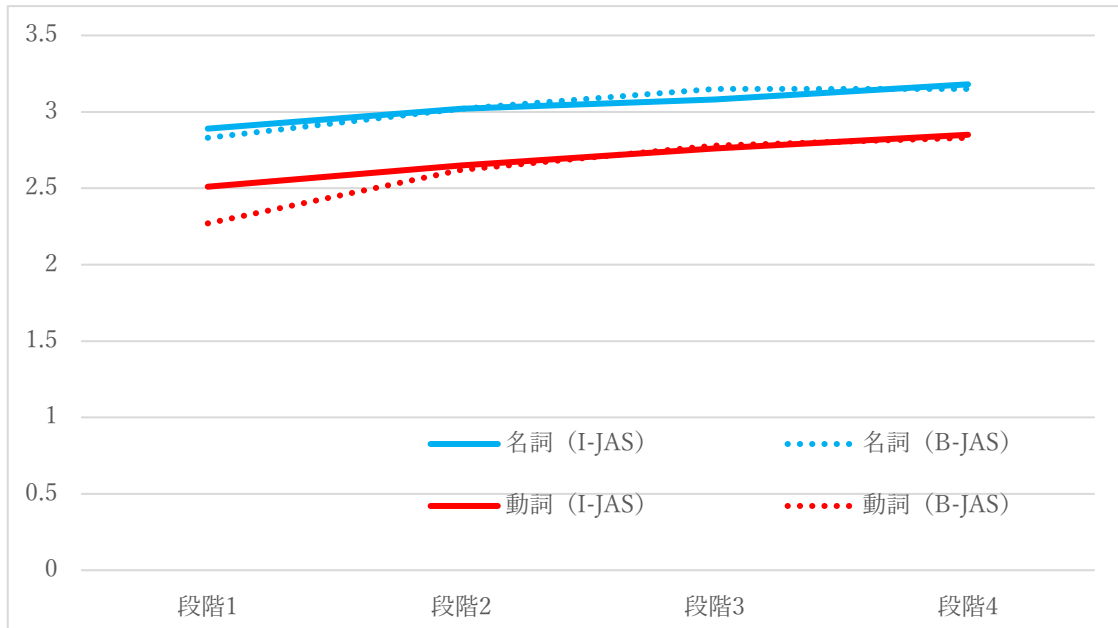


図 3 SW1 作文における名詞・動詞頻度の段階別変化

上記より、I-JAS、B-JAS とともに、(1)対数変換後の頻度で言うと、名詞はおよそ 3 前後、動詞はおよそ 2.5 前後になること、および、(2)段階が進むにつれて頻度はわずかに上昇するが、上昇幅は総じて小さいこと、が確認された。また、B-JAS の段階 1 を除くと、値も、I-JAS と B-JAS 間で非常に似通っていることが示された。

続いて、段階別に、1 人あたりの 3 種の助詞の使用頻度（トークン数を自然対数に変換）の変化を調査したところ、図 4 の結果を得た。

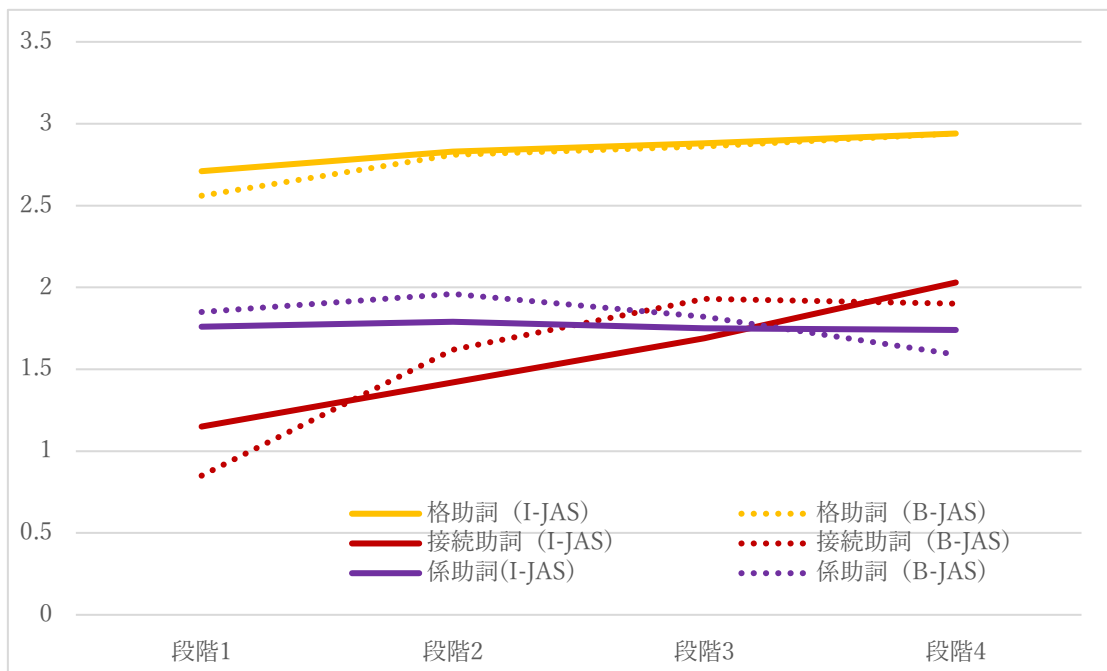


図 4 SW1 作文における助詞頻度の段階別変化

上記より、I-JAS、B-JAS ともに、(1)対数変化後の頻度で言うと、格助詞はおよそ 2.5～3、係助詞は 1.5～2、接続助詞は 1～2 となること、(2)格助詞と接続助詞は段階が進むにつれて増加すること、(3)係助詞頻度は減少すること、(4)増加幅は接続助詞が最も大きいこと、が確認された。ただ、接続助詞に関しては、I-JAS では増加幅がほぼ一定で、段階 1～4 にかけて線形的に増加するものの、B-JAS では増加幅が段階 1～2、段階 2～3、段階 3～4 になるにつれて縮小していくこともあわせて観察された。

以上、5 種類の品詞を取り上げて、その頻度変化のパターンを確認してきたが、総じて言えば、I-JAS、B-JAS 間で、高い一致度が示されたと結論できる。ただし、すべての値が同じように変化するわけではなく、接続助詞のように、増加という全体的トレンドは一致していても、増加の仕方に差が出る場合も存在することが示唆された。

4.3 RQ3 特徴語

I-JAS の 4 段階別データとそれらの総体、B-JAS の 4 段階別データとそれらの総体を比較し、各段階の特徴語を調査したところ、表 5 の結果を得た。表中、語彙素に付した数字は、左側が対数尤度比統計量、右側が相互情報量である。なお、対数尤度比統計量が 6.63 を超えていれば $\alpha=1\%$ 水準で、3.84 を超えていれば $\alpha=5\%$ 水準で、それぞれ頻度差が有意であることを示す。また、2.71 を超えていれば、 $\alpha=10\%$ 水準で、有意傾向にあることを示す。ただし、下記の中には、グループ内の 1 名ないし数名が顕著に使っているものも含まれている。

表 5 I-JAS と B-JAS における各段階の特徴語 ($\alpha=10\%$)

段階	I-JAS	B-JAS
1	食べ物 (8.4/1.1)、弁当 (7.6/2.2)、箱 (6.0/1.7)、悪い・御飯 (5.4/2.2)、ね (4.7/1.7)、思う (3.8/1.3)、郊外・驚く (3.7/1.5)、彼 (3.0/0.6)、ながら (2.9/1.3)、ます (2.9/0.3)	です (8.2/0.7)、有る (5.4/1.0)、へ (4.0/0.9)、は (4.0/0.3)、夫婦 (3.9/1.3)、ば・大変 (3.3/1.5)、分かる (3.3/1.2)、たい (3.2/1.0)、そして (3.1/0.9)、ケン (3.1/0.4)、マリ (2.8/0.4)、楽しい (2.8/0.8)
2	彼 (5.0/0.3)、等 (4.7/0.3)、時 (3.0/0.3)、家 (2.8/0.5)	食物 (3.6/1.2)、は (3.2/0.2)
3	(該当なし)	て (4.1/0.3)
4	て (8.8/0.5)、二人 (6.4/0.6)、居る (5.4/0.6)、共 (4.7/1.5)、たり・笑う・上・昼 (3.7/2.2)、内 (3.3/1.1)、気付く (3.1/0.9)、促 (2.7/1.3)、中 (2.7/0.6)	て (3.5/0.2)、来る (2.9/0.9)

まず、I-JAS について言うと、段階 1 では、基本語の不適切な使用が目につく。たとえば、ピクニックのために用意するサンドイッチを「食べ物」(犬は食べ物を食べて (CCH25))、

「弁当」(弁当のバスケット (CCH28))、「御飯」(ご飯を作りました (CCH30))と呼んだり、バスケットを「箱」と呼んだりする(犬が箱の中に入ります (CCH21))。また、サンドイッチが犬に食べられてなくなってしまったことや、それを知ってがっかりしたことを指して「悪い」と言ったり(食べ物は悪くになりました (CCM07) / 気持ちは悪くになりました (CCM14))、並行的な動作の進行を含意するはずの「ながら」を文脈に合致しない形で用いたりすることもある(彼らは地図を見ながら、犬がバスケットに入った (CCH63))。このほか、丁寧体語尾の「ます」、登場人物の心理的内面を推し量る「(〜と) 思う」、話し言葉的な終助詞「ね」、の使用も多い。

- (1) ケンはマリと楽しみにしていますけど、相談しながら、近く公園に行きます。(CCH21)
- (2) ケンさんとマリさんはとても怖いとおもいました。(CCH24)
- (3) 食べ物はなかったので、残念でしたね。(CCH025)

第2段階になると、2名の登場人物を言い換える「彼等」や、場所表現としての「家」、時間表現としての「〜する時」などが使用されるようになる。

- (4) 彼らはサンドイッチを作った後、地図を見ているとき、彼らが飼う犬はバスケットのサンドイッチを食べてしまいました。(CCH31)
- (5) ある日、ケンさんとマリさんが家でサンドイッチを作っていました。(CCH52)

そして、上級相当の第4段階になると、数詞によって登場人物を代名詞的に言い換える「二人」、アスペクト形式(接続助詞「て」+補助動詞「居る」)、列挙表現「たり」のほか、意味の比喩的拡張を伴う抽象的场所表現としての「上」「内」「中」などが特徴的に使われるようになる。

- (6) 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。(CCM35)
- (7) ですが二人が地図を見ている隙に、家で飼っていた犬がバスケットの中に入り込みました。(CCM51)
- (8) しかし、食べ物を望んでいた犬はバスケットの中に入って、食べ物を全部食べました上に、そのままかごの中に寝てしまいました。(CCH16)
- (9) 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。(CCM35)

以上、I-JASの段階別特徴語を概観すると、段階1~4にかけて、(1)主として基本名詞に関わる誤用の減少、(2)終助詞や丁寧体を用いた話し言葉的な表現の減少、(3)人物を言い換える代名詞表現の多様化、(4)アスペクト形式(ている)の獲得、(5)比喩的表現の獲得、といった大まかな方向性が見て取れる。

では、これと同じ傾向がB-JASでも見られるのであろうか。B-JASの特徴語について言うと、段階1では、丁寧体語尾の「です」や存在動詞の「有る」(仕事がありませんでした (CCB007))、取り立ての係助詞「は」や方向を含意する格助詞「へ」(犬はバスケットの中へ跳びました (CCB001))、仮定表現「ば」(大変と言えば大変ですけど (CCB007))、意思・願望を表す助動詞「たい」(犬はサンドイッチが食べたいでした (CCB011))、接続詞「そして」(そして、ケンとマリは地図を見ました (CCB006))などが特徴的に使用される。このうち、丁寧体助動詞と過去助動詞が連結した「でした」については、各種助動詞やイ形容詞など、本来は結合できない語群の後接事例が多く認められる(したいでした・分からないでした・したことはないでした (CCB003) / 知らないでした (CCB004) / 楽しいでした (CCB006) / 驚嘆でした・面白いでした (CCB007) / 悲しいでした (CCB013) など)。

なお、「美しいでした」のような結合について、教育出版株式会社編集局 (n.d.) は、中学校国語教員向けの解説において、誤用とはしていないものの、「美しかったです」という言い方のほうが一般的であるとしている。

段階 2～3 では、前述の取り立ての「は」に加え、バスケット内のサンドイッチを指して不適切に使われる「食物」(バスケットの中の食物 (CCB014) /食物は食べ終わりました (CCB005)) や、接続助詞の「て」を含むテイル形 (待っていました (CCB006) /相談していました (CCB012)) などが特徴的に用いられる。また、上級相当の段階 4 になると、引き続き接続助詞の「て」が多用され、とくに、テクル形が使用される (おなかもすいてきました (CB016) /ワンちゃんは出てきました (CCB007))。

以上を整理すると、B-JAS の段階 1～4 においては、(1)主として機能語的語彙に関わる誤用の減少、(2)取立て表現の増加、(3)アスペクト形式 (テイル・テイク) の獲得、といった方向性が認められる。

ここで、I-JAS の特徴語と比較すると、段階 1 については、丁寧体語尾の使用には一定の連関性が認められるが、単語のレベルでは一致しているものはない。また、段階 2～4 についても、テイル・テイク形の構成成分としての接続助詞の「て」を除くと、個別語レベルでの重複は見られない。I-JAS データの観察から演繹された 5 つの発達傾向のうち、B-JAS において明確に再現されたものは、段階 4 におけるアスペクト形式の獲得だけで、そのほかについてははっきりした一致は認められなかった。RQ1 および RQ2 で見たように、総語数や品詞頻度といったマクロ的なレベルでは I-JAS と B-JAS の調査結果は高い一致度を示したわけであるが、個別語のレベルになると、両コーパスの調査結果の重なりは限定的なものとなることが示唆された。

4.4 RQ4 段階分類

学習者 8 群と、10 品詞の頻度に対して対応分析を実施したところ、第 1 次元の寄与率が 78.7%、第 2 次元の寄与率が 8.1% で、上位 2 次元で全体の 8 割以上が説明された。また、学習者 8 群と、30 動詞の頻度に対して対応分析を実施したところ、第 1 次元の寄与率が 38.6%、第 2 次元の寄与率が 21.7% で、上位 2 次元で全体の 6 割以上が説明された。第 1 次元を横軸、第 2 次元を縦軸と散布図を作成したところ、図 5-6 の結果を得た。

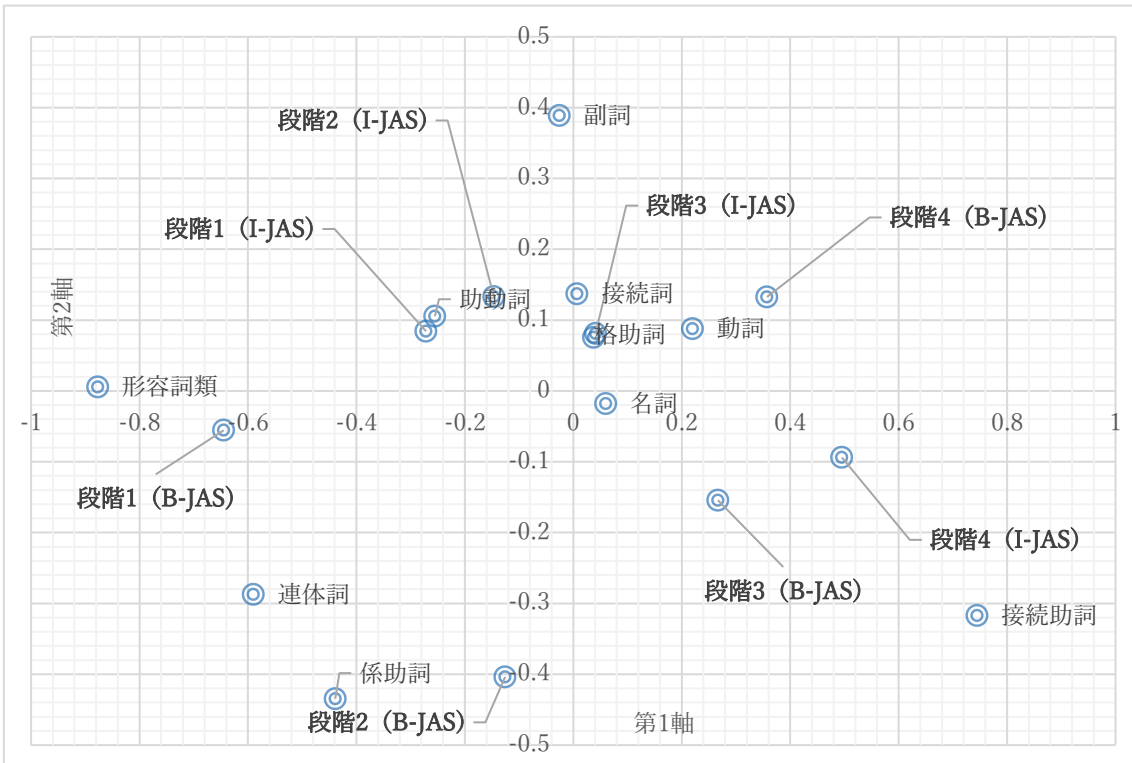


図 5 主要品詞と学習者 8 群の関係性

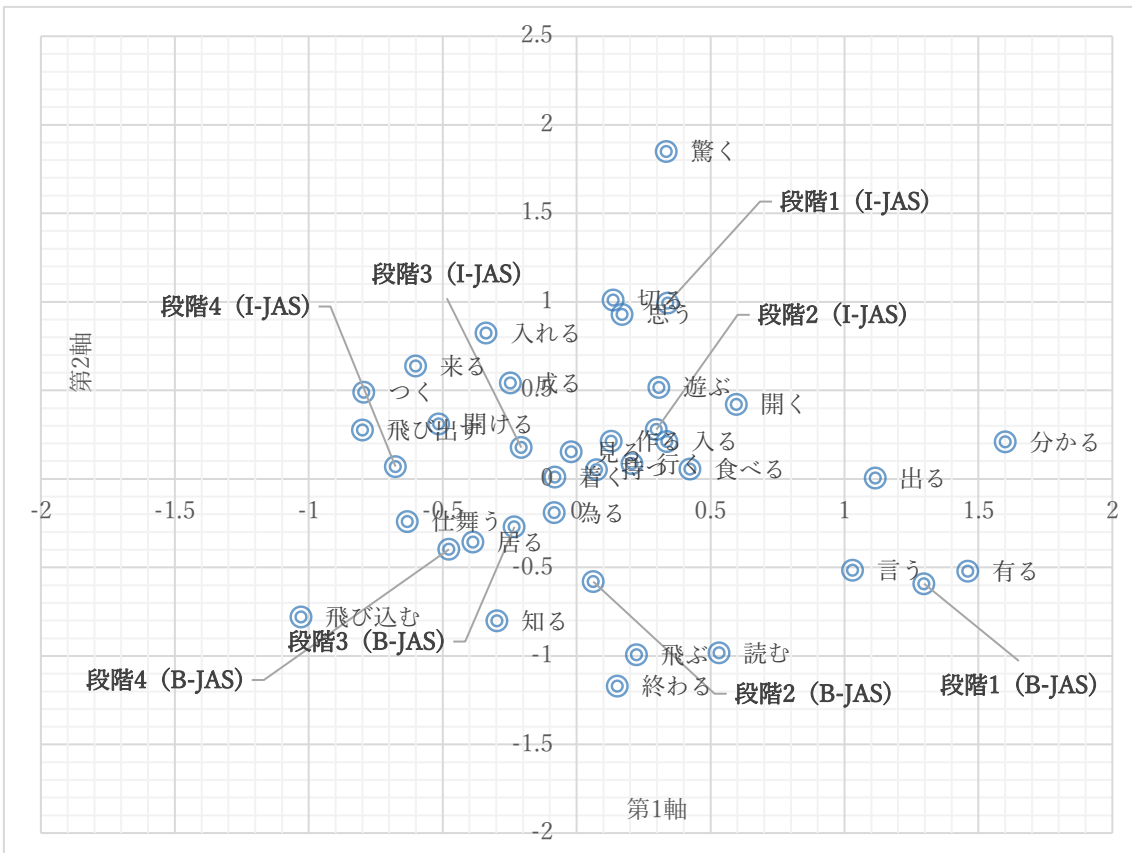


図 6 主要動詞と学習者 8 群の関係性

図 5-6 に示すように、10 品詞、30 動詞のいずれを手掛かりとした場合であっても、I-JAS、B-JAS とともに、次元 1 (横軸) 上で、段階 1 から段階 4 へ、同じ順序での布置が確認された (図 5 と図 6 で向きは逆転しているが、対応分析で得られた布置図における方向は相対的なものである)。このことは、品詞タイプ、あるいは、特定品詞に含まれる個別語セットというレベルで集約を行った場合、I-JAS であっても B-JAS であっても段階の差異が同じように現れることを示す。

しかしながら、I-JAS と B-JAS がすべてにおいて同じ結果を示すわけではない。実際、次元 2 (縦軸) 上で、I-JAS と B-JAS は上下に完全に分離されており (なお、品詞分類における段階 4 については、他の段階と異なり、上下の入れ替わりが見られる)、2 つのコーパスは異なる品詞ないし動詞タイプによって特徴づけられている。表 6 は、I-JAS と B-JAS における各段階の近傍に布置された特徴品詞・特徴動詞を整理したものである。

表 6 対応分析における I-JAS、B-JAS の各段階の近傍に位置する品詞・動詞タイプ

	品詞		動詞タイプ	
	I-JAS	B-JAS	I-JAS	B-JAS
1	形容詞	連体詞	開く、食べる、入る	有る、言う、読む
2	助動詞	係助詞	驚く、遊ぶ、行く、思う	飛ぶ、終わる
3	格助詞、接続詞、副詞	名詞	見る、着く、成る、入れる、開ける	為る、知る
4	接続助詞	動詞	来る、つく、飛び出す	居る、仕舞う、飛び込む

表 6 が示すのは、品詞についても、具体的な動詞タイプについても、I-JAS と B-JAS から演繹される学習者の L2 日本語の発達パターンが、一部の重複 (たとえば、段階 4 におけるアスペクト形式のテクルやテイル、また、複合動詞の使用) を除くと、それぞれ、かなり異なった形で見えることである。このことは、横断データか縦断データかのいずれか一方のみを根拠として L2 習得の議論を行おうとする場合、見落とされてしまう要素がありうることを示唆する。

5. まとめ

本稿では、横断コーパスである I-JAS と縦断コーパスである B-JAS の SW1 タスクにおける CLJ 産出をデータとして、習熟度・学習期間に基づく 4 つの段階別に、横断データと縦断データの同一性を検証してきた。

まず、RQ1 (総語数) については、I-JAS、B-JAS とともに、語数は 80~100 語に収まり、段階ごとに総語数は増加するものの、増加幅は段階があがるにつれて減少する、という同一のパターンを示すことが確認された。

また、RQ2（品詞頻度）については、I-JAS、B-JASともに、名詞・動詞・格助詞・接続助詞ともに段階が進むにつれて増加する一方、係助詞頻度は減少する、という共通のパターンを示すことが確認された。

次に、RQ3（特徴語）については、I-JASからは、基本名詞の誤用減少→口語的表現減少→代名詞表現多様化→アスペクト形式獲得／比喩表現獲得、といった発達パターンが見られるのに対し、B-JASからは、機能語的語彙の誤用減少→取立て表現増加→アスペクト形式獲得、といったパターンが観察され、アスペクト形式（テイル、テイク）の獲得を除くと、2種類のコーパスから導かれる段階別特徴は必ずしも一致しないことが示唆された。

最後に、RQ4（段階分類）については、主要品詞頻度と主要動詞頻度を手掛かりとした統計的分類により、I-JAS、B-JASともに、4段階の順列性は確認されたものの、品詞・動詞から見るL2習得プロセスは、一部の例外（上級段階におけるアスペクト形式と複合動詞の獲得）を除くと、コーパスによって異なる形で提示されることが示唆された。

以上、4つの研究設問の検討を通して、総語数や主要品詞別頻度の増減といったマクロ的観点については横断・縦断データともにほぼ同一の結果が得られるが、個々の語の使用に踏み込んだミクロ的観点については異なるパターンが析出されることが確認された。このうち、後者の知見は、とくに計量的視点からの学習者コーパス研究の実践において、横断データと縦断データをうまく統合して使用することの必要性を示唆するものと言えるだろう。

コーパスに限らず、第2言語習得研究全般における横断・縦断データの関係について、Gass, Behney, and Plonsky (2020) は、2種類のデータがそれぞれ長所・短所を持ち、かつ、それらが相補的な関係であることを強調している。横断データを用いた疑似縦断分析は、多くの学習者から得られる知見を一般化しやすい一方、群を対象とした分析を行うため、個々の学習者の詳細な背景を議論しにくい。また、習熟度の比較で発達が観察できるとする仮説の妥当性も完全には証明されていない。これに対し、縦断データを用いた分析は個々の学習者背景を加味した議論が可能であるが、サンプルが少ないため、分析はケーススタディ的・記述的・質的・状況説明（ナラティブ）的になりがちで、得られた知見の一般化は行いにくい。また、産出データが当該学習者のL2知識の全体を反映しているかどうか曖昧である（pp. 19-22）。Gassらの指摘をふまえれば、2種のコーパスの統合的使用のメリットはいっそう明白となるだろう。

もっとも、本研究には、さらに検討すべき点も多い。ここでは2点に絞って言及する。1点目は、今回の調査で示唆されたI-JASとB-JASの差異が、横断・縦断という学習者コーパスの収集方法の違い（のみ）に還元できるかどうかはいまだ断言しがたいということである。今回は、学習者の母語とタスクタイプをそろえることで、データ収集方法以外の要因を可能な限り統制して比較を行ったわけであるが、B-JASのデータは特定の1大学で集めたもので、当該大学における教材や教授法といった特殊要因が学習者の言語使用に影響している可能性も否定できない。

2点目は、今回は、I-JAS、B-JASともに、習熟度ないし学習期間で区別された群を対象として分析を行い、得られた結果を比較したわけであるが、そもそも、群特性が、当該群に

所属する個体の特性に完全に一致するかどうかは実のところははっきりしない。図 7 は、B-JAS の 17 名の学習者による 1 年次から 4 年次までの 4 回 (Y1-Y4) の SW1 タスクの総語数の変化をプロットしたものである。全体として平準化すれば、段階 1~段階 4 の間で漸増というパターンが現れるわけであるが、総語数といったごく基本的な観点ですら個体差が非常に大きいことに気づく。同じことは、I-JAS の習熟度別の 4 群にもあてはまる。このように高度に可変的な個人データを扱おうとすれば、2000 年代以降の応用言語学で注目を集めている複雑性動的システム理論 (complex dynamic system theory) などの知見を援用した新しい分析の枠組みが求められる (Hiver, Al-Hoorie, & Evans, 2022)。

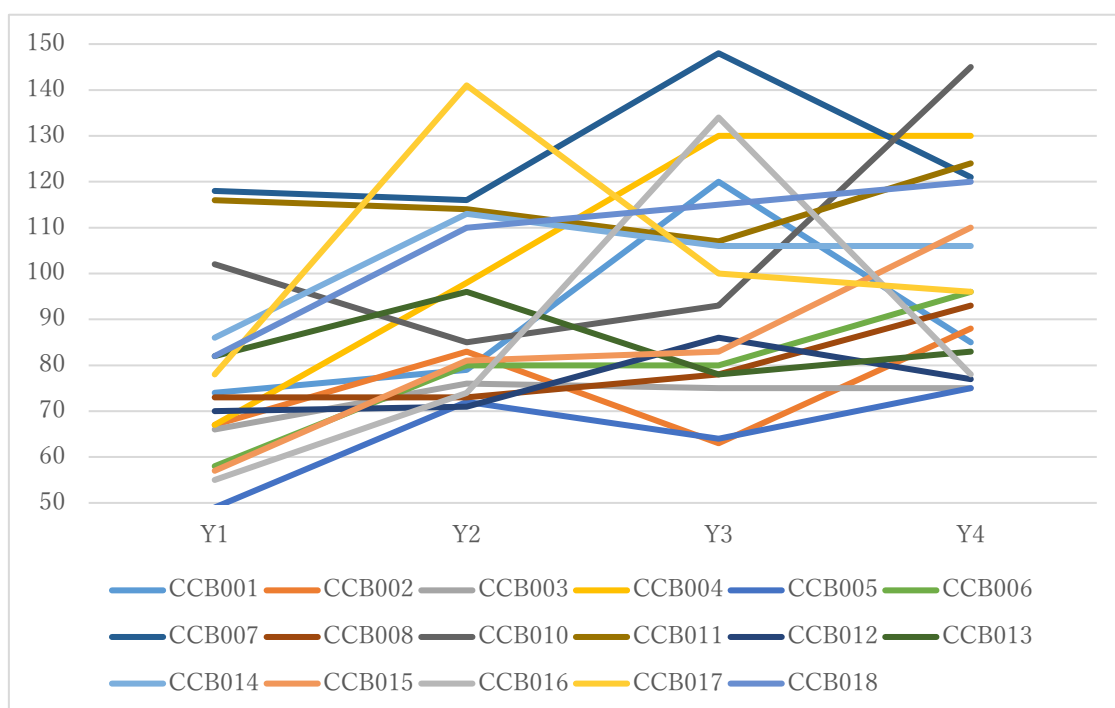


図 7 B-JAS の 17 名の参加者の SW1 総語数の変化

一般に、コーパス研究に関して想定される問題の多くは、データ量を拡張することで解消しうる。しかし、すでに内外で 100 億語を超えるコーパスが数多く構築されている母語話者コーパスと比べると、学習者コーパスのサイズははるかに小さい。この点をふまえると、学習者コーパスの研究者は、どのコーパスを使う場合であっても、その制約と限界についてこれまで以上に意識的であることが求められよう。多くの学習者コーパスが利用可能になってきた現在、任意の学習者コーパスから得られた結果が、別のコーパス、とくに、横断と縦断のように、本質的な性質を異にするコーパスにおいて再現されるかどうかを確認するプロセスは、学習者コーパス研究の信頼性を高める上で、ますます重要になってくると思われる。

とくに、I-JAS を用いて CLJ の習得研究を行おうとする場合には、B-JAS との照合が強く推奨される。仮に同じ結果が出れば、研究者は、得られた知見の妥当性をより明確に主張することができる。一方、異なる結果が出れば、群データから個別データに帰り、それぞれ

のコーパスに付属するフェイスシートを用いて個々の学習者の背景を検討し、必要に応じて、データを組み替えて、再分析にかけることができる。そもそも、I-JAS と B-JAS のように、同一デザインで横断・縦断的にデータを集めるという試みは、世界の学習者コーパス研究を見てもほとんど過去に例がない。2つのコーパスの革新性と、学習者コーパス研究史における意義を強調して本稿を閉じたい。

注

本研究は、国立国語研究所のプロジェクトによる成果である「多言語母語の日本語学習者横断コーパス (I-JAS)」および「北京日本語学習者縦断コーパス (B-JAS)」を利用して行われたものである。

文 献

- Gass, S. M., Behney, J., & Plonsky, L. (2020). *Second language acquisition: Introductory course*. (5th ed.). Routledge.
- Granger, S. (2002). A bird's eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson, S. (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). John Benjamins.
- Hiver, P., Al-Hoorie, A. H., & Evans, R. (2022). Complex dynamic systems theory in learning: A scoping review of 25 years of research. *Studies in Second Language Acquisition*, 44(4), 913–941.
- Ishikawa, S. (2017). Learners' acquisition and use of L2 Japanese vocabulary: Influence of L1 backgrounds and L2 proficiency levels: A learner corpus-based analysis. 『第二言語としての日本語の習得研究』 20, 10–27.
- Johnson, K., & Johnson, H. (Eds.). (1999). *Encyclopedic dictionary of applied linguistics*. Blackwell Publishing.
- Meunier, F. (2015). Developmental patterns in learner corpora. In Granger, S., Gilquin, G., & Meunier, F. (Eds.), *The Cambridge handbook of learner corpus research* (pp. 379–400). Cambridge University Press.
- Myles, F. (2015). Second language acquisition theory and learner corpus research. In Granger, S., Gilquin, G., & Meunier, F. (Eds.), *The Cambridge handbook of learner corpus research* (pp. 309–331). Cambridge University Press.
- Ployhart, R. E., & Vandenberg, R. J. (2010). Longitudinal research: The theory, design, and analysis of change. *Journal of Management*, 36(1), 94–120.
- Vyatkina, N., & Cunningham, D. J. (2015). Learner corpora and pragmatics. In Granger, S., Gilquin, G., & Meunier, F. (Eds.), *The Cambridge handbook of learner corpus research* (pp. 281–305). Cambridge University Press.
- 石川慎一郎 (2018a). 「L2 日本語語彙の習得プロセスについて—LARP at SCU コーパスを用いた台湾人学習者による日本語語彙運用の時系列分析—」『統計数理研究所共同研究リポート』 400, 1–18.

- 石川慎一郎 (2018b). 「中国語母語の日本語学習者の発話における使用語彙の変遷：発達段階の差と個体の差をめぐって」『国立国語研究所第4回学習者コーパス・ワークショップ & シンポジウム「第2言語習得における語彙の役割」予稿集』 62-76.
- 石川慎一郎 (2020). 「日本語・中国語・韓国語・英語母語話者の日本語発話における形容詞使用実態—I-JASに基づく調査—」『国立国語研究所日本語学習者コーパス「I-JAS」完成記念シンポジウム予稿集』 27-34.
- 石川慎一郎 (2021a). 「絵描写作文課題における L2 日本語学習者の動詞使用と習熟度の関係—I-JAS の SW1 課題データの計量的概観—」『統計数理研究所共同研究レポート』 444, 1-22.
- 石川慎一郎 (2021b). 「韓国学習者の日本語動詞獲得モデル：学習者総体モデルとの比較—『多言語母語の日本語学習者横断コーパス』の絵描写作文を用いた検証—」『日本語教育研究』(韓国日語教育学会) 56, 37-54.
- 教育出版株式会社編集局 (n.d.). 「ことばのてびき Q30 : 『美しいです』『大きいです』という言い方は正しいか」 <https://www.kyoiku-shuppan.co.jp/textbook/chuu/kokugo/guidanceq030-00.html>
- 迫田久美子(2020). 「I-JAS 誕生の経緯」迫田久美子・石川慎一郎・李在鎬(編著)『日本語学習者コーパス I-JAS 入門：研究・教育にどう使うか』(pp.2-13). くろしお出版
- 迫田久美子・石川慎一郎・李在鎬 (編著). (2020). 『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.
- 趙麗雯 (2015). 「学習者コーパスに見られる『テイナイ』の使用順序—縦断的・横断的観点から」『日本語／日本語教育研究』 6, 79-96.
- 李在鎬・小林典子・今井新悟・酒井たか子・迫田久美子 (2015). 「テスト分析に基づく『SPOT』と『J-CAT』の比較」『第二言語としての日本語の習得研究』 18, 53-69.