



日本語能力試験(JLPT)のN4～N1 の読解テキストの語彙特徴 : JLPT 読解テキストコーパスの分析

石川, 慎一郎

(Citation)

統計数理研究所共同研究リポート, 481:119-138

(Issue Date)

2025-03-25

(Resource Type)

departmental bulletin paper

(Version)

Version of Record

(JaLCD0I)

<https://doi.org/10.24546/0100493764>

(URL)

<https://hdl.handle.net/20.500.14094/0100493764>



日本語能力試験(JLPT)の N4～N1 の読解テキストの語彙特徴

— JLPT 読解テキストコーパスの分析 —

石川 慎一郎(神戸大学)

iskwshin@gmail.com

Lexical Features of the N4-N1 Reading Texts in the JLPT

— An Analysis of the JLPT Reading Text Corpus —

ISHIKAWA, Shin'ichiro (Kobe University)

概要

新しい外国人就労制度の整備に伴い、日本語能力試験(JLPT)の初中級レベル(N4 や N3)の合格に新たな意味づけがなされつつある。しかし、**JLPT** の試験問題や出題語彙基準が非公開になっていることもあり、入門から上級まで(N5～N1)、とくに、大学進学希望者から就労継続希望者の直接の学習目標となる初級から上級まで(N4～N1)を通貫したテキストの言語特性の実証的研究は必ずしも十分になされていないのが現状である。そこで本稿は、**JLPT** の公式問題集に掲載された N4～N1 の読解セクションのテキストを小規模コーパス化し、レベルごとの語彙特性を明らかにすべく、計量的な語彙解析を試みた。その結果、(1)N4 から N1 にかけてトークン数は 4.5 倍、タイプ数は 3.5 倍まで増加するが語彙多様性は不変であること、(2)高頻度語使用頻度に基づく、4つのレベルは N1・N2、N3、N4 の 3つに区分され、とくに N4 とそれ以外の間に境界があること、(3)主要品詞別にみると、多くの場合、初中級(N3～N4)と上級(N1～N2)の 2つに区分され、前者は並列・起点・終点を表す助詞、丁寧形助動詞、過去助動詞、日常名詞、丁寧形動詞などによって、後者は、引用・限定・仮定を表す助詞、受け身・否定を表す助動詞、形式名詞(もの、こと等)、抽象名詞、動詞過去形、動詞を含む固定句(という、よって等)によって特徴づけられることが示された。得られた結果は、就労希望者を対象とする日本語指導に有益であるのみならず、テキストの難度や平易さをめぐる言語学的な議論にも一定の示唆を有する。

キーワード

日本語能力試験、レベル別語彙量、特徴語、クラスター分析、対応分析

1. はじめに

1.1 新しい就労制度と日本語教育

2024 年 3 月に「出入国管理及び難民認定法及び外国人の技能実習の適正な実施及び技能実習生の保護に関する法律の一部を改正する法律」が公布され、2027 年 6 月までの施行が決ま

った。同法は、多くの問題が指摘されていた従前の「技能実習」を「育成就労」に置き換えるとともに、エントリーレベルの育成就労から、就業中は在留期間を無期とする特定技能 2 号に至るまでの道筋と、段階ごとの要求資格と待遇を明示化した制度設計となっている。政府は、「キャリアアップの道筋を明確化」するとともに、「労働者としての権利性の向上」を目指した新制度により、日本が、国際的な人材獲得競争の中にあって、海外の就労希望者から「選ばれる国」となる展望を示している(厚生労働省, 2024)。

この新就労制度は、日本語教育にも少なからぬ影響を及ぼすことが予想されている。というのも、表 1 に示すように、就労段階のステップアップ条件として、A1(初級)、A2(準中級)、B1(中級)といった欧州共通言語参照枠(CEFR)に基づく日本語レベルが明示的に組み込まれており、各段階にあることの挙証手段の例として、国際交流基金と日本国際教育支援協会が実施する「日本語能力試験」(以下、JLPT)の N5、N4、N3 レベルが表示されているからである。

表 1
新外国人就労制度における就労段階と必要な日本語能力

段階	必要な日本語能力水準	必要な業務能力	在留期間・帯同
育成就労	A1(JLPT の N5 等)以上、または同等レベルの日本語講習の受講。※ A1～A2(JLPT の N5～N4 等)の範囲で分野ごとに設定された日本語能力を取得すれば転籍が可能。	技能検定基礎級等	3 年(不合格の場合は最長 1 年延長可)
特定技能 1 号	A2(JLPT の N4 等)以上	相当程度の知識又は経験を必要とする 技能(技能検定 3 級 ／特定技能 1 号評価試験)	通算 5 年上限(例外 規定あり)
特定技能 2 号	B1(JLPT の N3 等)以上	熟練した技能(特定 技能 2 号評価試験)	制限なし+家族帯 同が可

注:法務省(2024, 2025)の内容を筆者が整理した。CEFR と JLPT レベルの対応は暫定版で、正式には、2025 年 7 月に決定される見込みである。

もちろん、日本語の能力テストには様々な種類があり、就労者の日本語力が A2 レベルに到達しているか否かを CBT(computer-based test)方式で診断する「国際交流基金日本語基礎テスト(JFT-Basic)」なども開発されているわけだが、試験としての知名度や普及度を考えると、一定数の就労者が JLPT レベルの取得を目指すのではないかと予想される。

1.2 JLPT

JLPT は 1984 年に始まった。年間受験者数(応募者数)は、1984 年時点では 7 千人程度であったが、1997 年に 10 万人、2007 年に 50 万人、2018 年に 100 万を超え、新型コロナ禍で一時的に減少したものの、2023 年に過去最高の 126 万人となり、名実ともに、日本語の習熟度試験の代表格となっている。

JLPT は 2010 年に大幅な改訂がなされ、表層的な言語知識だけでなく、それらを使って実際の言語課題(読解・聴解)を遂行する「言語コミュニケーション能力」のレベルを測定するという新しい理念が導入された。また、従前の全 4 レベル(4 級～1 級)における 3 級と 2 級の間に N3 が加わって全 5 レベル(N5～N1)に変更されたほか、素点による処理が等価スコアによる処理に変わり、さらに、詳細な出題基準や語彙リストなどが非公開となった。改訂の理念と改訂後の JLPT の概要を示した「新しい日本語能力試験ガイドブック」(国際交流基金・日本国際教育支援協会, 2009) (以下、ガイドブック)によると、N5～N1 の各レベルが前提とする能力は以下のようなものとなっている(p. 8)。

表 2

JLPT における N1～N5 の「認定の目安」

レベル	認定の目安
N1	(現実の生活の)幅広い場面で使われる日本語を理解することができる。
N2	(現実の生活の)日常的な場面で使われる日本語の理解に加え、より幅広い場面で使われる日本語をある程度理解することができる。
N3	(現実の生活の)日常的な場面で使われる日本語をある程度理解することができる。
N4	(教室内で学ぶ)基本的な日本語を理解することができる。
N5	(教室内で学ぶ)基本的な日本語をある程度理解することができる。

注：丸ガッコ内の内容は、JLPT サイト内の「4 つの特徴」(<https://www.jlpt.jp/about/points.html>)の記述を根拠に筆者が追記した。

JLPT の 5 つのレベルのうち、従前より、日本の大学に留学して学修を行うには JLPT の N2 以上が必要とされてきた(出入国在留管理庁, n.d.)。また、大学院レベルでは入学時に N1 を最低要件とするところも多い(例:神戸大学, 2024)。こうしたことから、留学希望者向けの日本語指導では N1 や N2 の合格を目指す指導が提供されてきたわけであるが、就労制度の改変に伴い、N4 や N3 の合格をターゲットとした就労(希望)者向け日本語指導というニーズが増加しつつある。表 1 で明らかなように、N4 ないし N3 に合格できなければ、就労を中止して帰国せざるを得なくなるわけで、N4 や N3 は就労者にとってきわめて high-stakes なテストとなる。

このように考えてくると、N4～N1 の各段階で要求される日本語能力をそれぞれ具体的に把握することが重要になるわけだが、現状のレベル基準(表 2)は、言語の使用場面を 3 段階(幅広い場面／日常的な場面／教室内の基本的な場面)、理解度を 2 段階(理解できる／ある程度理解でき

る)に分割して両者を組み合わせているだけで、それぞれのレベルで実際にどのような能力を問うているのかはこれだけでは把握しにくい。

この点について、外国人向けに平易化した「やさしい日本語」を研究する吉開(2018)は、一般の日本人は、外国人向けの JLPT を受験できないため、『N4』の試験に合格すると、どんなことが理解でき、何を語れるのか。これは日本語教育関係者でなければ知りえない」現状にあるとした上で、2010 年の改訂で、JLPT の「語彙・文法・漢字などの出題基準が公開され」なくなったことが、「やさしい日本語」を含む「研究の障害」になっていると指摘している。「外国人に(最低でも)N4 相当の日本語習得を求める以上、そのレベルがどのようなものかを日本人が理解し、コミュニケーションをすり合わせることもまた必要」であるという氏の指摘は説得力を持つ。

1.3 JLPT が問う読解能力

すでに述べたように、JLPT は、言語知識をベースとして、読解と聴解の課題遂行力を問う設計となっている。以下では、就労者が自然環境からのインプットだけでは身につけにくいと思われる読解力に議論を絞り、また、新しい就労制度の中で合格が必須化されていない N5 を除く N4～N1 の 4 レベルに限定して、実際の設問を通して、各レベルで要求される読解能力の詳細を考えていきたい。

JLPT は、読解テキストの選定について、以下のような方針を示している。

…テキストによって読み方が変わるので…海外や国内の学習環境を考慮した適切な範囲で、多様なテキストを扱います。テキストの話題・内容は、学習に関するもの、生活の中で目に触れる実用的なもの、仕事に関するものなどを取り入れます。また、テキストの種類は、説明文、意見文、評論、エッセイなどの他に、生活場面で目にする連絡や案内、仕事で使われる文書などです。テキストの形式は、一般的な文章形式の他に、箇条書きや表の場合もあります。そして、テキストの長さも、レベルに応じて短文、中文、長文の区分けを行います(ガイドブック, p.38)。

この方針に沿い、各レベルの読解テキストや設問が決定される。以下は、4 つのレベルごとに、前提とされている読解能力(言語行動)と、改訂で新規に導入された「統合理解」と「情報検索」を含む 6 種類の「大問のねらい」をまとめたものである。

表 3

JLPT におけるレベル別の読解能力の目安

N1	N2	N3	N4
・幅広い話題について書かれた新聞の論説、評論など、論理的にやや	・幅広い話題について書かれた新聞や雑誌の記事・解説、平易な評論な	・日常的な話題について書かれた具体的な内容を表す文章を、読んで理	・基本的な語彙や漢字を使って書かれた日常生活の中でも身近な話題の

複雑な文章や抽象度の高い文章などを読んで、文章の構成や内容を理解することができる。・さまざまな話題の内容に深みのある読み物を読んで、話の流れや詳細な表現意図を理解することができる。	ど、論旨が明快な文章を読んで文章の内容を理解することができる。・一般的な話題に関する読み物を読んで、話の流れや表現意図を理解することができる。	解することができる。・新聞の見出しなどから情報の概要をつかむことができる。・日常的な場面で目にする難易度がやや高い文章は、言い換え表現が与えられれば、要旨を理解することができる。	文章を、読んで理解することができる。
--	---	---	--------------------

(ガイドブック, p.8)

表 4

JLPT におけるレベル別の「大問のねらい」

	N1	N2	N3	N4
内容理解 (短文)	生活・仕事などいろいろな話題も含め、説明文や指示文など 200 字程度のテキストを読んで、内容が理解できるかを問う。	生活・仕事などいろいろな話題も含め、説明文や指示文など 200 字程度のテキストを読んで、内容が理解できるかを問う。	生活・仕事などいろいろな話題も含め、説明文や指示文など 150～200 字程度の書き下ろしのテキストを読んで、内容が理解できるかを問う。	学習・生活・仕事に関連した話題・場面の、やさしく書き下ろした 100～200 字程度のテキストを読んで、内容が理解できるかを問う。
内容理解 (中文)	評論、解説、エッセイなど 500 字程度のテキストを読んで、因果関係や理由などが理解できるかを問う。	比較的平易な内容の評論、解説、エッセイなど 500 字程度のテキストを読んで、因果関係や理由、概要や筆者の考え方などが理解できるかを問う。	書き下ろした解説、エッセイなど 350 字程度のテキストを読んで、キーワードや因果関係などが理解できるかを問う。	日常的な話題・場面を題材にやさしく書き下ろした 450 字程度のテキストを読んで、内容が理解できるかを問う。

内容理解 (長文)	解説、エッセイ、小説など 1000 字程度のテキストを読んで、概要や筆者の考えなどが理解できるかを問う。	—	解説、エッセイ、手紙など 550 字程度のテキストを読んで、概要や論理の展開などが理解できるかを問う。	—
	複数のテキスト(合計 600 字程度)を読み比べて、比較・統合しながら理解できるかを問う。	比較的平易な内容の複数のテキスト(合計 600 字程度)を読み比べて、比較・統合しながら理解できるかを問う。	—	—
主張理解 (長文)	社説、評論など抽象性・論理性のある 1000 字程度のテキストを読んで、全体として伝えようとしている主張や意見がつかめるかを問う。	論理展開が比較的明快な評論など、900 字程度のテキストを読んで、全体として伝えようとしている主張や意見がつかめるかを問う。	—	—
情報検索	広告、パンフレット、情報誌、ビジネス文書などの情報素材(700 字程度)の中から必要な情報を探し出すことができるかを問う。	広告、パンフレット、情報誌、ビジネス文書などの情報素材(700 字程度)の中から必要な情報を探し出すことができるかを問う。	広告、パンフレットなどの書き下ろした情報素材(600 字程度)の中から必要な情報を探し出すことができるかを問う。	案内やお知らせなど書き下ろした 400 字程度の情報素材の中から必要な情報を探し出すことができるかを問う。

(ガイドブック, pp.20-23)

表 3 および表 4 を見ると、N4 から N1 へと段階が上がるにつれて、要求される読解力は、テキストの長さ(100 字→1000 字)、場面(学習面→生活・仕事などの各面)、内容の抽象度(具体的→抽象的)、テキストの加工度(平易な書き下ろし→自然なテキスト)、テキストから把握すべき情報(内容のみ→因果関係→因果関係と理由)、テキスト数(単一テキスト→複数テキストの統合理解)などの点でより高度化していくことがわかる。こうした変化は、読解対象(テキストの部分／全体)と読解スタイル(迅速に／注意深く)という 2 観点に基づく「四つの読み方」に対応するものとされる(ガイドブ

ック、p.34)。

しかし、具体的にどのような言語特性・語彙特性を持ったテキストの読解が各レベルで期待されているのかはなお曖昧さが残る。留学希望者対象の N1～N2 指導と、就労者対象の N3～N4 指導を統一的にとらえるという観点からは、N4～N1 を通貫した読解テキストの言語特性の実証的な比較研究が必要になると思われる。こうした問題意識を出発点として、本研究においては、JLPT の公式問題集に掲載された N4～N1 の読解セクションのテキストを小規模コーパス化し、レベルごとの言語特性を明らかにすべく、計量的な語彙解析を試みることにしたい。

2. 先行研究

本研究のテーマにする関連する研究として、まず、JLPT 語彙リストの改訂作業の詳細を報告した押尾他(2008)を見ておきたい。JLPT を実施する国際交流基金と日本国際教育支援協会は 2005 年に「日本語能力試験の改善に関する検討会」を発足させ、2010 年の改訂に向けた準備作業が部会別に行われた。このうち、出題基準分科会の漢字表・語彙表部会においては、(1)作成方針と選別方針の決定およびデータベースの調査と整備、(2)語彙の一次選抜、(3)語彙の二次選別と記述方法検討、(4)語彙の三次選別と初中級語彙の検討、というプロセスで作業が進められた。語彙の選別に関しては、データベース頻度で採否を決めるという基本方針のもと、成人日本語話者の語彙量をふまえて、教育経験者の視点や旧試験出題基準語彙表も参考にされた。語彙選別のための資料としては、『日本語の語彙特性』(三省堂, 1999/2003)、『現代新国語辞典(改訂 3 版)』(学研, 2000)、『出題基準改訂版』(凡人社, 2004)、『現代雑誌の語彙調査:1994 年発行 70 誌』(国立国語研究所, 2005)を一次資料としつつ、外来語・オノマトペ・話し言葉・慣用表現に関する副次的資料も活用された。話し言葉については、「男はつらいよ」のシナリオ(CASTEL/J 研究会, 1969-1995)、筑波大学話し言葉資料(筑波大学, 1997-2001)、『女性のことば／男性のことば 職場編』(ひつじ, 1999/2002)、「BTS による多言語話し言葉コーパス」(宇佐美, 2003) などが参照され、最終的に、10,000～18,000 語程度の選定が目指されたと報告されている。なお、前述のように、作成された語彙リストは公開されていない。

一方、一般公開されていた旧試験の語彙リストは、2010 年の JLPT 改訂後も活用されている。とくに、川村よし子氏らが開発したオンラインの「日本語読解学習支援システム リーディング・チュウ太」(川村・北村・保原, 1997)には、入力した「本文中に含まれるすべての単語が日本語能力試験(本稿著者注:改訂前の旧試験のこと)の出題基準に準拠してレベル分けされ」、旧試験で言う「4 級の語は黒、3 級は灰色、2 級は青、1 級は赤、級外(出題基準以外)の語は濃い赤で表示」という「語彙チェッカー」の機能があり(川村, 2009, p. 19)、研究や教育の場に大きな影響を与えた(北村, 2013)。ただし、旧試験の語彙基準については、「カセット」「マッチ」のような今ではあまり使われない語が数十語存在する一方、最近の外来語が不足しているなどの問題もあるとされる(松下, 2017)。

語彙リストに加え、問題そのものも非公開となったことから、2010 年改訂後の JLPT のテスト問題の分析研究はあまりない。例外として、中町・佐藤・西内・浅原・奥村(2022)は、JLPT の学習参

考書から訓練データを、『日本語能力試験公式問題集』(凡人社, 2012)の言語知識・読解セクションから評価用データ作成し、訓練データから開発した文単位の難易度推定システムを評価データに適用してシステムの精度を検証した。その結果、開発したシステムによって文の実際の難度をある程度推定できること、文難度の推定において漢字比率や文字数といった基本統計量や単語難易度が説明変数として有効であること、「です」「ます」などの丁寧形文末表現が初級テキストに出やすいこと、高難度の N1 テキストにも初級の決まり文句が出てくるため、N1 推定の精度は他より低くなること、などが明らかになった。

このほか、JLPT 以外の試験テキストから言語特性の抽出を試みた研究もある。林(2022)は、JLPT の N1 とほぼ同等レベルとされる、日本学生支援機構(JASSO)が運営する「日本留学試験」(EJU)の 2010 年から 2019 年までの「日本語」読解分野の本文部分(延べ語数約 92,000 語)をコーパス化して重要語彙を抽出し、既存の日本語学術語彙リスト JAWL(松下, 2011a; 松下, 2011b)と比較して特徴を分析した。その結果、品詞と語種構成は「現代日本語書き言葉均衡コーパス」の一般的な書き言葉の場合と類似していること、延べ語数の 3 割、異なり語数の 56%が JAWL で言う中級レベルに当たること、専門用語が形態素解析で一般語に分解されてしまうなどの原因もあり(例:「調べ学習」→「調べる」+「学習」)、文献の内容分野(人文系・社会系・理工系・生物医学系)に固有の語は少ないこと、などが明らかになった。また、品詞による選別、JAWL とリーディング・チュウ太を用いた語彙のレベル付けと基礎語彙の除外、表記揺れの統合とエラーデータ修正、文書頻度(レンジ)による語彙ランキングという過程を経て、とくに非漢字圏学習者への EJU 読解指導の支援を目的とした「EJU 読解語彙リスト」(中級レベル:883 語)と、指導前の確認に使用できる「EJU 読解初級語彙リスト」(初級レベル:223 語)が作成された。

以上の研究は、それぞれ価値ある知見を披歴しているが、JLPT の読解テキストに限り、N4～N1 の語彙特性の変化を実証的に論じた研究はほとんど存在せず、研究の必要性が残されていると言える。

3. リサーチデザインと手法

3.1 研究目的と RQ

本研究は、JLPT の N4～N1 の読解問題テキストを素材とする小規模コーパスを作成し、各レベルで出題される読解テキストの語彙量および語彙特性を概観することを目的とする。この目的に沿い、3 種の研究設問を設定する。

RQ1 レベルの変化によって、読解テキスト全体の語彙量はどのように変化するか？

RQ2 高頻度語全体に着目した場合、4 つのレベルはどのように区分されるか？

RQ3 名詞・動詞・助動詞・助詞に着目した場合、4 つのレベルはそれぞれどのように区分され、各区分はどのような語彙で特徴付けられるか？

3.2 データ

JLPT 実施団体が刊行する「日本語能力試験公式問題集」(凡人社, 2012: 以下 v1)および「日本語能力試験公式問題集第二集」(凡人社, 2018: 以下 v2)に掲載された N4~N1 の読解パートの問題テキスト(v1 は 37 種、v2 は 38 種、合計 75 種)を手作業で入力し、「日本語能力試験読解テキストコーパス」(JLPT Reading Texts Corpus)を構築した。出典表示、注釈、読み仮名、設問を示す下線数字などは入力対象外としている。レベル別の読解問題の構成は以下のとおりである。

表 5

読解問題の構成

	N1	N2	N3	N4
大問とテキスト数	大問 8 (3/4)*	大問 10 (5)	大問 4 (4)	大問 4 (4)
	大問 9 (3)	大問 11 (3)	大問 5 (2)	大問 5 (1)
	大問 10 (1)	大問 12 (2)	大問 6 (1)	大問 6 (1)
	大問 11 (2)	大問 13 (1)	大問 7 (1)	
	大問 12 (1)	大問 14 (1)		
	大問 13 (1)			
テキスト数計	23 (11+12)	24 (12 x 2)	16 (8 x 2)	12 (6 x 2)

注: N1 の大問 8 における問題テキスト数は v1 が 3、v2 は 4 である。

2 種の公式問題集のデータは別個に扱うこととし、8 種の統合ファイルを作成して分析対象とした (N1_v1、N1_v2、N2_v1、N2_v2、N3_v1、N3_v2、N4_v1、N4_v2)。

3.3 手法

本稿では、複数テキストから統合頻度表を自動生成する EJWFTG(石川, 2024a; 石川, 2024b) の V2.0 を用い、8 種の統合ファイルから語彙表を作成した。EJWFTG では、語彙表は、表層形(品詞なし)、表層形(品詞あり)、語彙素(品詞なし)、語彙素(品詞あり)の 4 モードで自動出力される。

まず、RQ1 (語彙量変化)については、4 レベル 8 種の統合ファイルごとに、トークン数(延べ語数)、タイプ数(語種数)、語彙多様性を調査した。ファイルごとのトークン数が大きく異なることから、語彙多様性は、type-token ratio(TTR)に代えて、タイプ数の自然対数をトークン数の自然対数で割る Herdan's C 値を使用する(石川, 2021, p. 144)。

次に、RQ2 (高頻度語に基づくレベル区分)については、全体での高頻度 50 語(表層形、品詞区別なし)をケース、8 種の統合ファイルを変数とする頻度表に対して階層クラスター分析を行い、変数側を分類する。距離は、相関係数に基づく(2-2r)の平方根で定義し、融合後の距離計算は Ward 法による。

表 6

全体頻度上位 50 語

クラスター分析に使用した 50 語(表層形)
の、に、て、は、を、が、で、と、た、も、し、こと、ない、な、ます、ある、いる、か、する、だ、から、いう、です、人、や、お、日、い、よう、それ、ば、もの、時、その、自分、この、まし、そう、れ、館、利用、者、なく、円、ください、図書、あり、的、さ、中

最後に、RQ3 (品詞別高頻度語に基づくレベル区分と区分特徴語)については、助詞・助動詞・名詞・動詞の 4 品詞の各々につき、8 種のファイルを第1アイテム、全体頻度の上位 20 語を第 2 アイテムとする分割表に対して、対応分析を実施する。第 2 アイテムに使ったのは以下の語である。

表 7

品詞別頻度上位 20 語

品詞	上位 20 語
助詞	の、て、は、に、を、が、と、で、も、か、から、や、ば、など、たり、まで、だけ、ながら、へ、より
助動詞	た、で、ます、な、だ、に、です、ない、まし、れ、ませ、ん、られ、だろう、たい、たら、なく、れる、られる、ず
名詞	こと、人、もの、時、自分、利用、円、図書、子ども、時間、人間、必要、本、とき、日、月、ため、ところ、中、世界
動詞	し、ある、いる、する、いう、い、ください、あり、なる、なっ、き、さ、できる、でき、よっ、考え、つい、なり、あっ、なら

対応分析とは、分割表の行列の相関を最大化する少数の次元(軸)を取り出す多変量解析手法で、通例、上位 2 次元を横軸・縦軸とする散布図上にアイテムカテゴリデータを布置し、データ間の相互の親疎関係を検討する。散布図に基づき、4 レベルの区分状況を整理し、各区分を特徴づける語彙の特定を行う。

4. 結果

4.1 RQ1 語彙量変化

レベルが上がれば読解で要求される語彙量も増加するかどうかを検証すべく、8 種の統合ファイル別の語彙量、また、v1、v2 の値を平均化した 4 レベル別の語彙量を調査したところ、以下の結果を得た。なお、タイプ数と語彙多様性については、表層形に基づくカウントと語彙素に基づくカウントを行い、前者を A、後者を B と記載する。

表 8

JLPT 読解テキストの語彙量(8 ファイル比較)

レベル	版	トークン数	タイプ数 A	タイプ数 B	語彙多様性 A	語彙多様性 B
N4	v1	661	239	217	0.843	0.828
	v2	786	271	238	0.840	0.821
N3	v1	1,574	492	440	0.842	0.827
	v2	1,590	476	414	0.836	0.817
N2	v1	2,581	748	661	0.842	0.827
	v2	2,567	711	615	0.836	0.818
N1	v1	3,185	886	792	0.841	0.827
	v2	3,385	883	776	0.835	0.819

表 9

JLPT 読解テキストの語彙量(4 レベル別平均値比較)

レベル	版	トークン数	タイプ数 A	タイプ数 B	語彙多様性 A	語彙多様性 B
N4	av	723.5	255	227.5	0.840	0.820
N3	av	1,582	484	427	0.840	0.820
N2	av	2,574	729.5	638	0.840	0.820
N1	av	3,285	884.5	784	0.840	0.820

N4～N1 における語彙量の変化の観察に先立ち、2 点を予備的に確認しておきたい。なお、タイプ数については、A と B で傾向に差はないため、ここでは B に限って概観する。

1 点目は、各レベルにおける v1 と v2 の語彙量の差についてである。仮に、同レベルの読解テキストであっても、試験の実施回によって語彙量に大きな差があるとするならば、これらを資料とする本分析の妥当性は棄損されることになる。この点に関して、両版におけるトークン数の差分と、少ないほうのトークン数の比率を調べると、ずれの度合いは N4 では 19%とやや多いが、N3 では 1%、N2 では 0.2%、N1 では 6.2%となり、差は総じて小さいことが確認された。また、タイプ数についても、ずれの度合いは N4 で 9.7%、N3 で 6.3%、N2 で 7.5%、N1 で 2.1%となり、版の間での顕著な差はみられなかった。

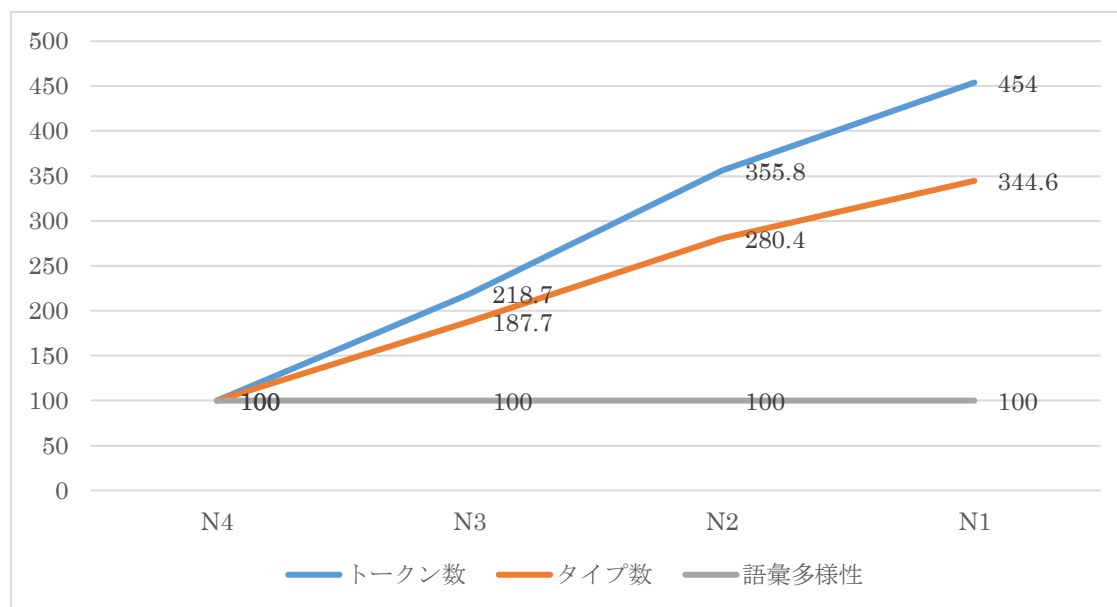
2 点目は、就労者の日本語習得上の新たな目標となる N4 と N3 の読解テキストの語種数についてである。タイプ数は、N4 であれば 230 種、N3 であれば 430 種程度となっている。現行の N4 に相当する旧 JLPT の 3 級では、「基本的な文法・漢字(300 字程度)・語彙(1500 語程度)を習得し、…簡単な文章が読み書きできる能力」が要求されていたが、1 回の試験に限った場合、出題される語彙の範囲は制約的であることが確認された。

以上の確認をふまえ、N4～N1 にかけての語彙量の変化を概観しておきたい。N4 の v1、v2 の平均値を 100%とみなして、トークン数、タイプ数 B、語彙多様性指標 B が N3、N2、N1(いずれも

v1, v2 の平均値)でどのように変化するか整理したところ、以下の結果を得た。

図 1

N4～N1 間の語彙量の変化



トークン数とタイプ数はいずれも線形的な上昇を示しており、トークン数は N3、N2、N1 とレベルが上がるにつれ、N4 の約 2.2 倍、3.6 倍、4.5 倍まで増加する。タイプ数は、トークン数に比べると上昇幅が緩やかであるが、約 1.9 倍、2.8 倍、3.4 倍まで増加する。レベル間での上昇の割合はほぼ同等であり、読解テキストに使用される語彙が高度にコントロールされていることを傍証する。

一方、語彙多様性指標はほぼ不変である。このことは、レベル上昇による読解テキストの語彙量の増加が、語彙多様性の増加、すなわち、テキストにおける異語使用度の上昇に起因するものというより、テキストの総量(トークン数)の増加の影響であることを示唆する。

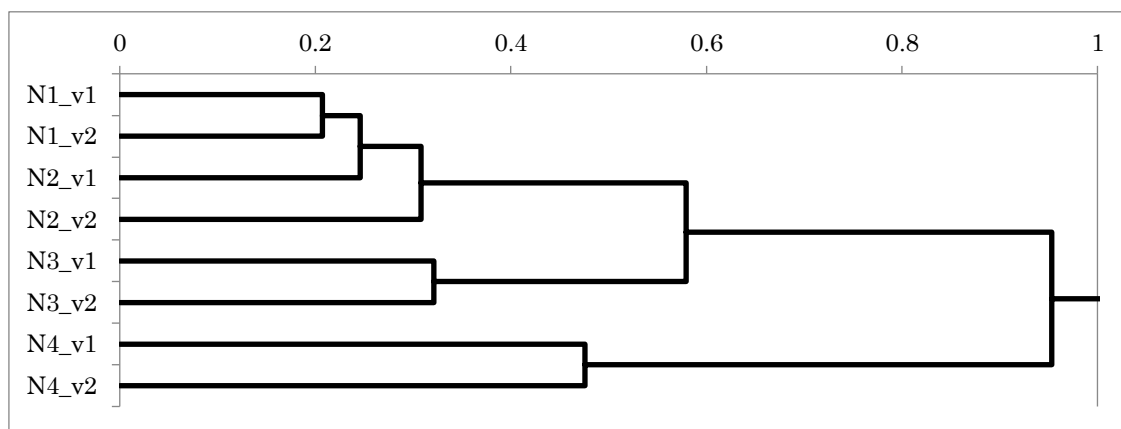
なお、大きな差ではないが、N4～N2 の増加度に比べると、N2～N1 の増加度はわずかに下がっているように見える。これは、N1 に近づくにつれ、語彙量に関しても、ある種の天井効果が生じている可能性を示す。

4.2 RQ2 高頻度語に基づくレベル区分

全体における高頻度語に着目した場合、JLPT の 4 つのレベルがどのように区分されるかを検証すべく、全体における頻度上位 50 語をケース、8 種の統合ファイルを変数として変数クラスター分析を行ったところ、以下のデンドログラム(樹形図)を得た。

図 2

クラスター分析に基づくデンドログラム



デンドログラムの解釈では、通例、その状態が変化しない定常状態が最も長い箇所にカットイングポイントを設定してデータの分割を行う。この方針に従い、0.6～0.9の間にカットイングポイントを置くと、8種の統合ファイルはN4の2ファイルとそれ以外に分割される。次に定常状態が長いのは0.3～0.5で、ここで再分割を行うと、残りの6ファイルはN3の2ファイルとそれ以外に区分される。以上より、JLPTの4つのレベルの関係は以下のように整理できる。なお、破線の長さはレベル間の距離の相対的な大きさの違いを示す。

N4-----N3-----N2--N1

N2とN1間の距離は相対的に見て小さい。実際、今回の分析サンプルでも、N2_v1は、N2_v2ではなく、N1側と先に融合しており、N1とN2の読解テキスト中の上位語彙使用には一定のオーバーラップも示唆されるところである。なお、N2とN1間の距離が小さいこと背景としては、前述の天井効果が影響している可能性も考えられる。

なお、4レベル中、特定技能2号の要件となるN3は、本来は「N1、N2とN4、N5の『橋渡し』のレベル」であるが(ガイドブック, p.61)、今回の語彙分析では、N3は、N4側よりも、むしろN1～N2側に近いという結果になった。就労者は口頭のインプットには触れる機会が多いが、書き言葉を読む機会は必ずしも十分でなく、N3の合格には語彙学習の面で相応の対策が必要になるかもしれない。

4.3 RQ3 品詞別高頻度語に基づくレベル区分と区分特徴語

主要品詞(名詞・動詞・助動詞・助詞)別に見た場合、JLPTの4つのレベルがどのように区分されるか、また、各区分がどのような語彙で特徴づけられるかを検証すべく、8種の統合ファイルを第1アイテム、品詞別の上位20語を第2アイテムとして対応分析を行ったところ、以下の散布図を得た。

図 3

助詞散布図

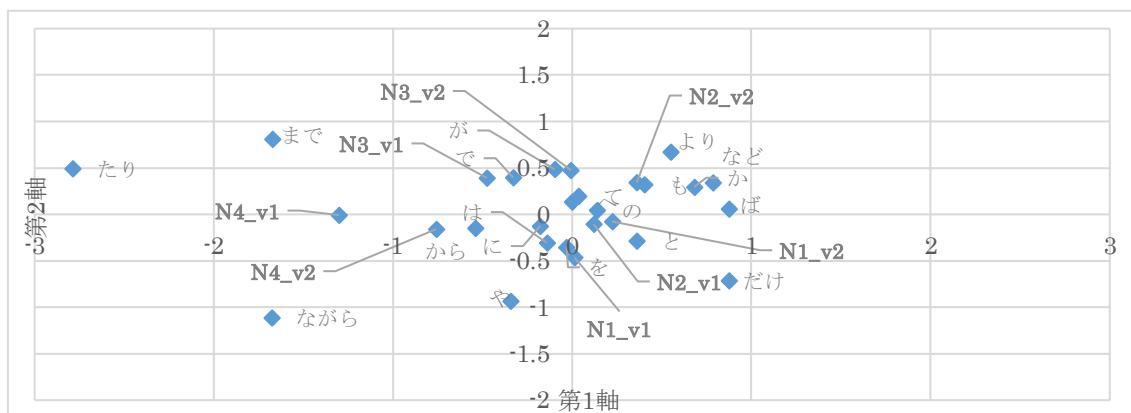


図 4

助動詞散布図

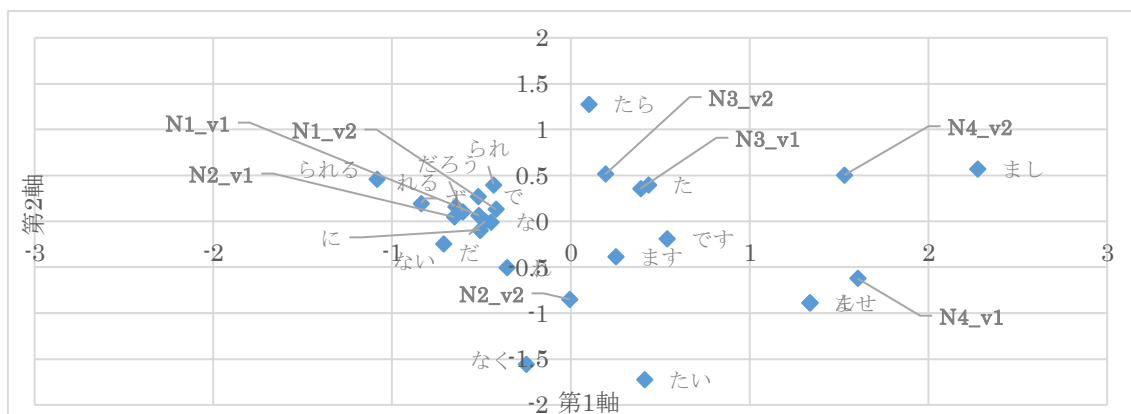


図 5

名詞散布図

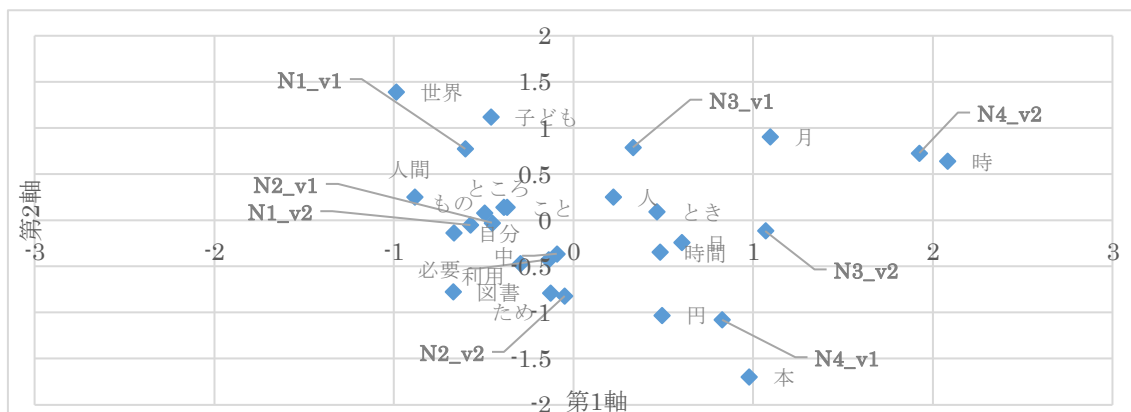
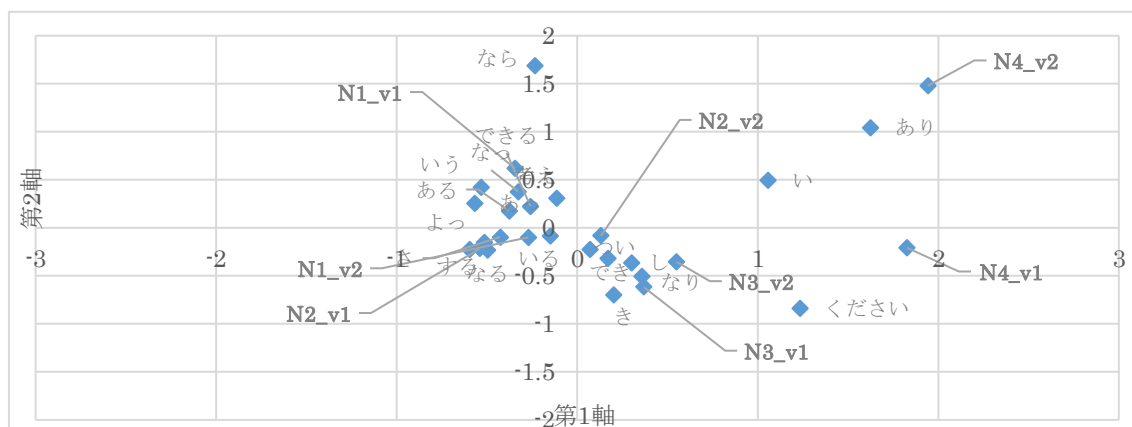


図 6

動詞散布図



はじめに、得られた散布図の信頼性について確認しておきたい。第1次元(横軸)と第2次元(縦軸)の説明力は、名詞では42.1%と18.4%(累計60.5%)、動詞では65.7%と12.1%(累計77.8%)、助動詞では48.3%と17.4%(累計65.7%)、助詞では40.1%と27.4%(累計67.5%)となり、4品詞ともに、分散の約40~65%が第1次元のみに集約され、60~80%が上位2次元に集約されていた。分散の集約度に関して言えば、名詞が低く、動詞が高い。名詞はテキスト内容に直接的に影響されやすいことから集約度が低下するが、全体として見れば、得られた散布図(特に第1次元)は元の分散の相当部分を説明するものになっていることが確認された。

続いて、4レベルの区分について概観する。散布図の解釈で最も重要になるのは、説明力の大きい横軸上で、左右いずれの側に布置されているかである(なお、散布図内の左右の関係は相対的なもので、左右いずれであるかに絶対的な意味はない)。今回の結果を見ると、1件の例外(動詞におけるN2_v2)を除き、すべてにおいて、N1とN2が左右の一方に、N3とN4が他方にかたまっていることが確認された。RQ2で見たように、全体での高頻度語に基づく区分では、N3はN2~N1側に近かったわけであるが、品詞を分けて観察を行うと、N3はN4側に入り、全体は、N4~N3とN2~N1に大きく分割されることとなった。なお、縦軸上での上下の区分に関しては、レベルによるまとまりは認められず、縦軸は各品詞における語彙の意味的な違いを表しているものと判断される。

上記の観察は、初中級(N3~N4)と上級(N1~N2)という大まかな区分ごとに特徴語が存在することを示唆する。そこで、第1次元スコアを根拠として、すべての語彙を、初中級特徴語、中立語、上級特徴語の3種に区分した。対応分析では原点付近に布置されたものは中立的な位置づけと解釈するのが通例であるため、以下では、原点0をはさみ、原点近傍の±0.3未満の範囲に入るものを中立語と分類している。

4品詞のうち、助詞については、第1次元のプラス側に上級が、マイナス側に初中級が布置されている。一方、助動詞・名詞・動詞の3品詞においてはマイナス側に上級が、プラス側に初中級が布置されている。そこで助詞を除く3品詞の第1次元スコアについて事前に-1をかけて調整を行

った。これにより、下表において、スコアの低さは初中級テキスト性を、スコアの高さは上級テキスト性を示すこととなる。

表 10

初中級読解テキストおよび上級読解テキストの特徴語

初中級特徴語			中立語		上級特徴語	
助詞	たり	-2.787	に	-0.178	と	0.360
	ながら	-1.677	は	-0.141	も	0.402
	まで	-1.672	が	-0.098	より	0.550
	から	-0.540	を	-0.031	か	0.682
	や	-0.344	へ	-0.001	など	0.785
	で	-0.330	て	0.034	ば	0.874
助動詞			の	0.139	だけ	0.876
	まし	-2.275	ます	-0.251	れ	0.357
	ませ	-1.336	たら	-0.101	られ	0.432
	ん	-1.336	なく	0.251	な	0.447
	です	-0.539			だ	0.481
	た	-0.433			に	0.507
	たい	-0.412			で	0.518
					だろう	0.606
					れる	0.645
					ない	0.713
					ず	0.838
					られる	1.086
名詞	時	-2.080	人	-0.219	こと	0.370
	月	-1.094	中	0.093	ところ	0.390
	本	-0.977	ため	0.128	子ども	0.461
	日	-0.601	必要	0.140	もの	0.495
	円	-0.491	利用	0.295	自分	0.669
	時間	-0.480			図書	0.672
	とき	-0.463			人間	0.886
動詞					世界	0.988
	あり	-1.624	き	-0.200	なっ	0.328
	ください	-1.232	でき	-0.168	ある	0.379
	い	-1.056	つい	-0.071	さ	0.428
	なり	-0.356	考え	0.116	なる	0.498
	し	-0.301	いる	0.152	する	0.514
			なら	0.235	いう	0.533
			できる	0.259	あつ	0.569
					よつ	0.596

以下、実際の用例を確認しつつ、初中級と上級を特徴づける語彙の特徴を見ておこう。

まず、助詞に関して、初中級用の JLPT 読解テキストには、並列や同時進行を示す「や」(例: クラブ活動やサークル活動[N3_v2])、「たり」(例: よい本の紹介文を書いたり…絵本を読んだりします[N4_v1])、「ながら」(例: 川の生き物に実際に触ったりしながら楽しく学べます[N3_v2])、起点・終点を示す「から」と「まで」(例: 駅から…バス停まで[N4_v1])などが多い。一方、上級用の読解テキストになると、引用・例示の「と」(例: 論文という形式[N1_v1])や「～など」(例: 遊びなどの実体験[N1_v1])、追加を示す「～も」(例: 苦しんだこともある[N2_v1])や、限定を示す「～だけ」(例: 間近の目標を達成していだけで十分だ[N2_v2])、疑問等を示す「か」(例: どちらかと言えば[N2_v1])、仮定を示す「ば」(例: 悪い面もあればいい面もある[N2_v1])など、初中級の場合以上に多様な項目間の関係性が助詞によって表出されている。

助動詞に関して、初中級では、丁寧形助動詞「ます」の活用形である「まし」や「ませ+ん」(例: 修理を頼みました[N3_v1])、丁寧形文末表現の終始形である「です」(例: 明日の約束ですが…[N4_v1])などが多用される。このほか、過去の「た」(例: 音楽のようだと感じた…[N3_v1])、願望の「たい」(例: 息子はテレビが見たい、ゲームがしたいと…[N4_v1])なども多い。一方、上級では、受け身を表す「れ」や「られ」(例: 遺伝子で決定されている[N1_v1])、推量の「だろう」(例: 余裕があったのだろうか[N1_v2])、否定の「ない」や「ず」(例: 知らず知らずのうちに[N1_v2])が多用される。話者やその感情が前面に出た初中級と違い、上級では、話者がテキストの背面に回るとともに、話者のスタンスの微妙なニュアンスの違いが表出されている。

続いて、名詞に関して、初中級では、イベントのチラシなどの読解が求められるため、日時や費用に関して、「月」「日」「時」「時間」「とき」「円」などが多用される(例: 自転車を借りる人は 200 円かかります[N4_v1])。一方、上級になると、名詞の性質は大きく変容し、「こと」「ところ」「もの」など、動詞要素を名詞句に取り込む形式名詞が多用される(例: 何の意味も持たないことが多いのです[N1_v1]／理屈を超えたところに[N2_v1]／自分が書いたものを読む相手[N2_v2])。また、「自分」「人間」「世界」なども多く、初中級で多かった個人的で身近な話題が、社会的なレベルに引き上げられていることが示唆される(例: 自分の世界がどんどん広がっていく[N2_v2])。

最後に、動詞については、初中級では、「あり」「なり」「ください」など、丁寧形のものが多い(例: 机の上にこのメモがあります[N4_v2]／自分が嫌になります[N3_v2]／取りに来てください[N4_v2])。これに対し、上級になると、「ある」「なる」「する」といった基本動詞の終始形に加え(例: 正直であることを強要しますが[N1_v1]／出掛けるようになると[N2_v2]／狩りをするのができないが…[N1_v2])、「なっ(た)」「あっ(た)」などの過去形、また、同格の「(と)いう」や、手段を表す「よって」のような決まり文句の一部(例: 自分を観察しているということを…[N1_v2]／経験則によって裏付けられた物…[N1_v2])などが多く、動詞のテンスやアスペクト、また、構文内での機能の多様性が大きく増していることが示される。

以下は、これらの特徴を多く含む初中級および上級の読解テキストの実例である。上記の統計分析で検出された特徴語を四角で囲んで示す。

表 11

初中級読解テキストおよび上級読解テキストの例

初中級読解テキスト	上級読解テキスト
(1) あおぞら—日スポーツ教室…5 月29 日 …スポーツ・時間・場所・お金…サッカー・ 10:00～17:00・第2運動場・100 円…ルー ルや、やり方を先生が教えます。サッカーは …10:00 までに来てください。(N4_V1_Q6) (2) 本についての相談を受けたり、どこにどの 本を置くのがいいか考えたりします。それか ら、図書館の新聞によい本の紹介文を書い たり、小さな子どもたちに絵本を読んだりしま す。(N4_v1_Q4) (3) 子どものころのことを思い出してみてくださ い。雲が動物の形に見えたり、壁のしみや 汚れが顔に見えたり し たことはあり ませ ん か。(N3_v1_Q4)	(1) 作法は、思想と結びつきながら伝承され てきたものであつた。たとえば昔は、食事の 作法を厳しくしつけられた。食べ物を残すこ とはもちろんのこと、さわぎながら食事をする ことも、けっしてしてはいけなかった。それ は、食事は生命をいただくものだ、と いう厳 かな思想があつたからである。茶碗の中の 米だけをみても、人間はおそらく何万と い う生命をいただかなければならない。 (N1_v1_Q12) (2) …わたしたちの認識は、自分の生きてきた 時代や環境に大きく左右される。ある意味、 閉じ込められているといってもいい。認識で きる「世界」はきわめて限定的なのであり、時 代や環境の制約によつて、認識の鋳型がで きてしまうから、場合によつては、大きく歪め られた「世界」像しか見えなくなる こと もあ る。(N1_v1_Q9)

本節の分析で得られた特徴語リスト、および、特徴語を含むテキスト実例を見ることで、JLPT の「認定の目安」だけではつかみにくかった異なる難度レベルの読解テキストの言語特性の違いがかなり具体的に抽出できたことになる。

5. まとめ

以上、本稿は、JLPT の公式問題集の読解セクションのテキストをコーパス化したうえで、各種の計量調査を実施した。

まず、RQ1(語彙量変化)については、(1)N4 から N1 にかけてトークン数は 4.5 倍、タイプ数は 3.4 倍まで増加すること、(2)語彙多様性指標はほぼ不変であること、また、(3)同一レベルの異なるテスト回による語彙量の差はきわめて小さいこと、(4)就労者の新たな学修目標となる N4 では 230 種程度、N3 であれば 430 種程度の語種が読解テキストに出現することなどが確認された。

次に、RQ2(高頻度語に基づくレベル区分)については、全体上位 50 語に基づくクラスター分析を行った結果、(1)全体は N4 とそれ以外に分かれ、次いで N3 と N1～N2 に分かれること、(2) N1 と N2 間で重複が確認されること、(3)特定技能 2 号の要件となる N3 は N4 側でなくむしろ N1～

N2 側に近いことなどが明らかになった。

最後に、RQ3(品詞別高頻度語に基づくレベル区分と区分特徴語)については、品詞別上位 20 語を対象とした対応分析により、(1)4 つのレベルがおよそ初中級(N3～N4)と上級(N1～N2)に区分されること、(2) 初中級の読解テキストでは、並列・同時進行・起点・終点を示す助詞、丁寧形・過去・願望を示す助動詞、イベントチラシなどに多い日時・費用に関連する名詞、丁寧形の動詞などが多用されること、(4)上級の読解テキストでは、引用・例示・追加・限定・疑問・仮定を示す助詞、受け身・推量・否定を示す助動詞、形式名詞と抽象名詞、終始形動詞、過去形動詞、決まり文句の一部をなす動詞(という、よって)などが多用されることなどが明らかになった。

以上で得られた知見は、JLPT の N4～N1 の読解テキストの言語特性を通観する際の一つの視座となるものであると同時に、幅広いレベルの日本語学習者、とくに、新就労制度の導入によって、新たに N4 ないし N3 の合格を目指す必要が出てきた就労者のための日本語指導の在り方や、いわゆる「やさしい日本語」を含め、外国人にとっての平易な日本語の在りようを検討する際の基本的な資料となるものと言える。

もっとも、本研究で分析できたテキスト量は決して多いものではなく、分析データのさらなる拡張が不可欠であろう。JLPT の公式問題そのものは非公開であるため、今後は、(データとしての正当性には制約があるものの)各種の JLPT 予想問題集のデータなども追加しつつ、本研究で得られた知見の再現性の確認を行っていきたい。

謝辞

本研究は科学研究費(23K25338)による研究成果の一部である。

引用文献

- 林希和子(2022).「日本留学試験読解コーパスの作成と分析:非漢字圏高等教育機関進学希望者への支援を目的として」『専門日本語教育研究』24, 67-74.
- 石川慎一郎(2021).『ベーシックコーパス言語学 2 版』ひつじ書房.
- 石川慎一郎(2024a).『森を見ながら木を見る』コーパス研究の意義: 複数テキストから統合語彙頻度表を作成する EJWFTG の開発『統計数理研究所共同研究レポート』469, 95-122.
- 石川慎一郎(2024b).「English/Japanese Word Frequency Table Generator (EJWFTG)を用いた日本語統合語彙頻度表の作成と活用」『計量国語学』34(6), 421-431.
- 川村よし子(2009).『チュウ太の虎の巻』くろしお出版.
- 川村よし子・北村達也・保原麗(1997).「日本語読解学習支援システム リーディング・チュウ太」
<https://tinyurl.com/y6c9j2dh>
- 北村達也(2013).「日本語読解学習支援システム『リーディング・チュウ太』」『甲南大学紀要:知能情報学編』6(2), 243-253.
- 神戸大学(2024).「令和 7 年度神戸大学国際文化学研究科研究生募集要項」
<https://tinyurl.com/289n6kry>

- 国際交流基金・日本国際教育支援協会(2009).「新しい日本語能力試験ガイドブック」
<https://tinyurl.com/2ck9d96n>
- 国際交流基金・日本国際教育支援協会(2012).『日本語能力試験公式問題集』. 凡人社. (N1
～N5 の 5 分冊).
- 国際交流基金・日本国際教育支援協会(2018).『日本語能力試験公式問題集第二集』. 凡人社.
(N1～N5 の 5 分冊).
- 厚生労働省(2024).「改正法の概要(育成就労制度の創設等)」外国人介護人材の業務の在り方
に関する検討会第 6 回資料(2024 年 3 月 22 日). <https://tinyurl.com/28zl2z7z>
- 法務省(2024).「育成就労制度の概要」 <https://tinyurl.com/2b2or82z>
- 法務省(2025).「基本方針の概要(案)について」第 1 回特定技能制度及び育成就労制度の基本方
針及び分野別運用方針に関する有識者会議(令和 7 年 2 月 6 日)の資料 3-1.
<https://tinyurl.com/2cg974mm>
- 松下達彦(2011a).「日本語の学術共通語彙(アカデミック・ワード)の抽出と妥当性の検証」『2011
年度日本語教育学会春季大会予稿集』 244-249.
- 松下達彦(2011b).「日本語学術共通語彙リスト Ver. 1.01」 <https://tinyurl.com/26fm5usz>
- 松下達彦(2017).「日本語読解テキストのリライトの重要性とアプローチ:語彙的要素を中心に」『日
本言語文化研究会論集』13, 1-18.
- 中町礼文・佐藤敏紀・西内沙恵・浅原正幸・奥村学(2022).「日本語能力試験に基づく日本語文の
難易度推定」『言語処理学会 第 28 回年次大会 発表論文集』 658-663.
- 押尾和美・秋元美晴・武田明子・阿部洋子・高梨美穂・柳澤好昭・岩元隆一・石毛順子(2008).「新
しい日本語能力試験のための語彙表作成にむけて」『国際交流基金日本語教育紀要』4,
71-86.
- 出入国在留管理庁(n.d.).「高等教育機関等へ入学するための日本語能力について」
<https://tinyurl.com/29zb7guy>
- 吉開章(2018).「日本語能力試験の語彙リスト公開を:『N4 って何?』と言われる前に」日本語教育
情報プラットフォーム:にほんご ふらっと「時代のことば」. 2018 年 6 月 13 日付記事.
<https://tinyurl.com/27c3xsnj>

注:上記のオンライン資料については、いずれも 2025 年 3 月 1 日時点で閲覧・確認。