

大規模テストの読解問題作成過程への コーパス利用の可能性

李 在 鎬

要 旨

本稿では、大規模テストの問題作成におけるコーパスの利用可能性を考察した。考察においては言語テスト分野におけるコーパス利用の現状を紹介した後、試験問題の作成過程でコーパスを用いる利点について述べた。そして、読解の問題作成での利用を想定し、日本語能力試験の級区分に基づくコーパスデータの分析を試みた。分析においては、日本語能力試験の読解テキストを学習データ、『現代日本語書き言葉均衡コーパス』の収録データを評価データにし、判別分析を行った。この分析から『現代日本語書き言葉均衡コーパス』の収録データが1級から4級のどの級に相当するかを明らかにし、問題作成における利用範囲を示した。最後に、日本語教育におけるコーパス分析の視点としては、定量的アプローチだけでなく、具体的な用例を確認しながら質的に分析していくアプローチも必要かつ重要であることを指摘した。

【キーワード】日本語能力試験, 現代日本語書き言葉均衡コーパス, 多変量解析 (特に判別分析), 真正性

1. はじめに (背景と目的)

大規模なテキストデータの集合からなるコーパスは、言語に関わるあらゆる領域に新しい可能性をもたらす。自然言語処理の分野では、コーパスを使うことで効率よく言語現象を構造化することができ、従来の規則や制約に基づく研究モデルでは得られない高精度な言語処理が可能になった。また、記述言語学・理論言語学の分野では、コーパスを使うことで従来の内省に基づく研究成果の検証を行うなどして、新しいタイプの実証研究ができるようになった。さらには、応用言語学・外国語教育の分野では、コーパスを使うことで言語の運用実態、実際の使用状況に即した教育コンテンツの開発が個人単位でも実現できるようになった。こうした多分野における様々な方向での応用が可能になった背景には、コーパスが有する資料としての科学性、汎用性が関係しているが、本特集のテーマとなる言語テスト分野に関しても、近年、英語教育を中心にコーパスの利用可能性に関して、盛んに議論されている (Alderson 1996, Chappelle 2001, Huston 2002, Barker 2004, Rimmer 2006, Chodorow et al. 2010 参照)。とりわけテスト問題の作成における自然さの検討、テスト課題の真正性検討、さらには問題分析時の信頼性・妥当性の検証などで、有効な支援ツールになるものとして注目されている (He & Dai 2006 参照)。

本稿では、大規模日本語テストを支援するものとして、コーパスの位置づけおよびその利用可能性について考察する。次の順序で議論する。まず、2節では応用言語学における

コーパス利用の現状を紹介した後、言語テスト分野における活用の可能性および課題について述べる。次に3節と4節では、2節の議論を踏まえ、日本語能力試験の級区分に依拠して、コーパスデータを統計的に分析した結果を報告する。5節では、4節の結果に対する考察を行うと同時に、日本語テストおよび日本語教育におけるコーパス活用の可能性を検討したい。

2. 先行研究と問題提起

コーパスは、一般的には工学的な言語処理の分野で使用することを主目的に作られたもので、言語処理の分野では辞書に次いでもっとも基本的な言語資源であるとされている。言語処理において、辞書は言語の知識に関わる言語資源であるが、コーパスは言語が実際に書かれたり、話されたりした結果が記録されたものであるため、言語の使用実態に関わる言語資源であると位置づけられている。コーパスが持つ特徴的性質は以下のキーワードでもって表すことができる。

1. 機械可読性
2. 処理の柔軟性
3. データの大量性

まず、1の性質として、コーパスは一般にコンピューター上で読み取りができる形式として構築される。このことが2の性質に関係してくるが、コーパスはデータ形式においてはユーザーが様々な処理を加えることが可能なものになっており、構造的な「柔軟性」を持っている⁽¹⁾。そして、コーパスの最大の魅力と言えるのが、3に示した「大量性」である。この「大量性」こそが、個人単位では得ることのできない客観性・科学性を導いており、新たな研究開発への応用を可能にしてくれる。こうしたコーパスが持つ特性は、90年代後半から、言語研究の分野においても注目されるようになり、応用言語学や外国語教育学の分野でもコーパスに基づく研究方法が急速に普及した⁽²⁾。

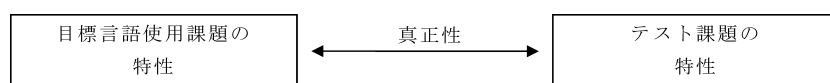
以下では、応用言語学におけるコーパス利用の現状について簡単に紹介した上で、言語テスト分野における利用の現状と課題について考えてみたい。まず、コーパス言語学の分野でもっとも研究が進んでいる言語とされる英語の場合、早い時期からデータベースの構築が行われ、80年代後半には基本的な研究インフラが整った。具体的には1964年に作られたBrownコーパス（アメリカ英語の書き言葉のコーパス）ができてから、80年代以降の急激なコンピューターテクノロジーの発展や90年代以降のウェブの進化に後押しされ、大規模化と多様化が急速に進んだと言われている。大規模化の方向としては、British National Corpus (BNC) や Bank of English (BoE) など、億単位のコーパスが次々と公開されたことが挙げられる。多様化の方向としては、「特定目的のための英語」(English for Specific Purposes; EPS) に対する関心の広がりから、医学や工学、司法などの特定分野に特化したコーパスが公開されていること、誤用分析や中間言語分析の流れをくみ、学習者言語を集めた言語習得コーパスなども公開されていることなどが挙げられる⁽³⁾。こうしたコーパスデータの整備による恩恵をもっとも直接的に受けたのは、外国語学習のための辞

書編纂の分野である。というのは、コーパスが持つ「大量性」は、当該言語の代表的な言語使用の実態を強く反映しているため、見出し語の決定に始まって、共起文脈の探索、語義の抽出、用例の選定に至るあらゆるプロセスにおいて、重要な参考資料になる⁽⁴⁾。具体的にはコーパスを活用することによって、語や句の出現頻度やコロケーション情報などを短時間で求めることができる。さらには用法に関する新しい事実も比較的容易に発見できるため、辞書記述の精緻化や内容の改訂においてもコーパスの力は絶大と言える。

さて、言語テスト分野におけるコーパス活用はどのような状況にあるだろうか。この分野において比較的古い研究として、Alderson (1996) があるが、それによれば、コーパスは問題の作成から採点に至るまで、あらゆるプロセスにおいて活用されるべきと主張しており、方法論構築の必要性についても指摘していた。そして、2000年以降、BNCやBoEといった大規模コーパスの公開や応用言語学分野でのコーパス研究の普及に足並みを揃える形で、ETS⁽⁵⁾やCambridge ESOL⁽⁶⁾などの世界的なテスト機関で、独自にコーパスを構築し、そのコーパスを基に作題をし、問題を評価し始めている。具体的な例としては、Barker (2006) ではCambridge ESOLにおけるコーパス利用の現状について紹介している。例えば、Cambridge ESOLでは、作文データで構成された「Cambridge Learner Corpus」という独自の誤用タグ付き学習者コーパスを構築し、内部資料として活用していることを報告している⁽⁷⁾。こうした学習者コーパスを利用することで、テスト開発者は、学習者の生きた言語使用を直接観察することができる。そして、テスト受験者がどのような表現において困難を感じているかを具体的に把握することができる。具体的な利用例としては、文法や語彙の問題作成時に、どのような錯乱枝を使うのが良いかについて事前シミュレーションができる。また、別の試みとして、BNCのような大規模な母語話者コーパスを作題時に活用していることも報告されている。BNCのような母語話者のコーパスは次のような活用方法が考えられている。まず、問題作成の段階においては、1) ターゲット項目の真正な使用場面を確認するため、2) 自然な用法を反映した言語表現を生成するため、3) ターゲット項目のもっとも汎用的な形式を見つけるため（英語の例でいえば、単数形と複数形のどちらで使うことが多い名詞かなど）にコーパスを活用できると述べている。次に、問題編集や検討時には、1) 自然な共起や出現頻度をチェックするため、2) 発話場面との整合性をチェックするため、3) 語彙の意味を形成する百科事典知識など、辞書にはない潜在的な意味を確認するため、コーパスを活用できると指摘している (Ibid: 2)。

言語テスト分野におけるコーパスの活用可能性を考える上で、深く関わりを持っているものとして、真正性 (authenticity) という概念がある。

真正性とは、図1の通り、ある言語テスト課題の特性と目標言語使用課題の特性の一致の度合いを表す概念として定義されている。言い換えれば、テスト課題が実生活 (real-



(バックマン, L. F., パーマー, A. S.(2000:27)

図1 目標言語使用課題とテスト課題

life) の言語使用にどの程度近いかを表す概念ということになる。こうした概念が重要視されるようになった背景には、まず学習者は自らが学習している目標言語を代表するテキストに触れながら、学習していく必要があるという問題意識があった。このことが言語テストにおいても、テストのためのテストにならないためには、常に考慮されるべきものであるという認識が広まった。とりわけ Bachman & Palmer (1996) では、真正性はテスト課題と得点解釈の一般化の領域 (domain of generalization; 目標言語使用領域の中の課題の集合体) とを関連づけるものであると指摘しており、テスト開発における最重要課題であると位置づけている⁽⁸⁾。具体的には、テスト受験者がテスト以外の場面で、目標言語を使用する状況を想定し、そこで出会う可能性が高い連の言語使用課題を記述し、テスト開発に反映させることが重要であると言える。そのためコミュニケーションテストやプロフィシエンシーを測るテストにおいて、真正性の高いテストを開発することは、テストとしての信頼性そのものに関わることであり、様々な形で工夫がなされる。例えば、スピーキングのテストであれば、実際に会話をさせて、それを評価したり、ライティングの試験であれば、実際に文章を書かせて、それを評価する、さらにはノートテーキングの試験であれば、実際の講義ノートをとらせてその結果を評価することなども考えられる。しかし、実生活での言語使用の多様性の問題、テスト実施上の環境的問題、さらには採点などの実務的な問題などを勘案した場合、テスト条件として完全な真正性を確保することは難しい。現実的なところでは、種々のテスト資料が目標言語の使用状況をどの程度、リアルに反映しているのかということが大きなポイントになる。一般的には、テストのために作られた素材ではなく、テスト以外の目的のために作成された言語素材、例えば新聞記事やドラマのシナリオなど、実際に話されたもの、実際に書かれたものを効率よく活用することで、真正性の高いコミュニケーションテストという目標に近づけることができるとされている。とはいっても、受験者の能力や学習環境の問題があるため、生の言語素材をそのままの形で使用することは難しいという現状もある。

では、日本語のテストを作成する際、テスト資料や素材の真正性をいかに確保すべきだろうか。この問題を実務レベルで捉えた場合、素材の真正性に対する客観的な指標が強く求められる。この要請に対して、コーパスは書き言葉が中心になるという問題はあるが、一つの解答を示してくれる。具体的な例として、「京都大学格フレーム⁽⁹⁾」のデータを検索することで、特定の述語がどのような助詞や名詞類と共起しているかを大量のデータから確認することができるが、そこで得られる出現頻度や共起頻度は表現の適格性や自然さを判断する上で、有効な指標になりうる。また、『BTS による多言語話し言葉コーパス』⁽¹⁰⁾ などでは (話者間の) 複数の関係がバランスよく設定されているため、言語表現と発話場面の整合性を検討することができる。さらに、進んだ利用方法として、『現代日本語書き言葉均衡コーパス (以下、BCCWJ と呼ぶ)』の場合、2009 年度のモニター版において 3000 万語規模の書籍データが含まれているが、このような大量のコーパスデータに対して、多少の編集を行って、読解問題のテキストとして活用することも十分可能である。しかし、これら、いずれの活用方法においても一つ大きな障壁になることとして、コーパスデータが持つ生の言語素材としての粗さを、言語テストという文脈においてどう位置づけるかという問題がある。というのは、コーパスデータは生の言語使用をリアルに

反映しているというメリットがある反面、日本語学習者や受験者の能力ということから考えてみた場合、無条件に容認できるものでもない。なぜならいわゆる日本語教育的レベル感ということについて、コーパスは全くのブラックボックスだからである。

以上を踏まえ、本研究ではコーパスデータに対し、日本語能力試験に基づく級区分を試みた。具体的には、これまでの日本語学的研究・計量国語学的研究で広く使われたコーパスデータを対象に、判別分析という統計的手法を使って日本語能力試験の1級から4級の(読解問題のテキストとして)どの級に相当するかを特定する計算機実験を行った。このような分析を行うことで、第一の期待として、どの級のテスト問題の作成や検討に、どのようなコーパスを参考資料として活用すべきかについての指針が得られる。第二の期待として、読解問題のテキストとして、コーパスデータの利用可能性についても何らかの示唆が得られるであろう。

3. データと方法

コーパス実験は判別分析という統計的手法を用いて行った。判別分析とは、事前に与えられているケースが異なるグループに属することが明らかな場合、新しいケースがどちらのグループに入るのかを判別する基準を得るための手法である。アルゴリズム的には二段階で構成されている。

- ・【第一段階】：すでにどのグループに属しているかが明確なケースの集合によって判別の基準が作成される。
- ・【第二段階】：どのグループに属するかが不明なケースを【第一段階】で作成した判別基準を用いて、属するであろうグループに判別、分類する。

本研究では、【第一段階】のデータ(以下、学習データと呼ぶ)として、級区分(属するグループ)が明確な日本語能力試験の読解テキストを使用した。具体的には、1984年から2007年まで、日本語能力試験において実際に使用された277個の読解テキストデータを使用した。次に、【第二段階】のデータ(以下、評価データと呼ぶ)としては、BCCWJの2009年度モニター版に含まれている「書籍」、「Yahoo!知恵袋」、「国会会議録」のデータを使用した⁽¹¹⁾。調査対象となるコーパスの選択における全体的な方針として、学習者に密接な言語使用文脈という観点も取り入れつつ、様々な文体のテキストを対象に調査すること、日本語学や言語処理などの関連分野でも広く利用されているテキストを対象に調査することを心がけた。具体的な選択の理由としては、次のようなことが挙げられる。まず、「書籍」コーパスは、書き言葉のもっとも代表的な形式であり、BCCWJの中でもっとも中心を占めるデータであることを踏まえ、調査対象にした。次に、「Yahoo!知恵袋」は、著作権フリーのデータであるなど、利便性の良さから、近年注目されていること、さらにウェブデータに見られる特徴として、書き言葉の特徴と話し言葉の特徴をともに反映していることを踏まえ、調査対象にした。最後に、「国会会議録」は松田(2008)のように日本語学的研究においてよく利用されていること、李(他)(2009)が示しているように、話し言葉を記録したものでありながら、新聞のような硬い文体のテキストとも類似

表1 実験データの規模

区分	データ名	文字数	形態素数	漢字数
学習データ	1級読解	75,802	48,584	21,172
	2級読解	66,598	50,927	15,614
	3級読解	15,728	13,062	2,491
	4級読解	9,225	7,331	609
評価データ	書籍 (BK)	174,948	119,496	47,907
	国会会議録 (国会)	412,353	238,599	122,269
	Yahoo! 知恵袋 (Yahoo)	93,856	59,988	20,981

点が多いことを踏まえ、調査対象にした。

さて、実際の調査では、技術的な部分で二つの先決課題があった。一つ目として、学習データの中に長さが異なるデータが混在すること、二つ目として、評価データの中に長さが異なるデータが混在することである。一つ目の課題は、日本語能力試験の読解テキストの場合、700～1000字前後で構成される長文読解と120～300字前後で構成される短文読解があり、受験級や実施年度によって長さが異なっている⁽¹²⁾。このように不均一なデータセットのまま統計分析を行った場合、誤った判別基準が作られる可能性に加え、ケース間の平等な比較ができないという問題が発生する。二つ目の課題は、BCCWJの場合においても「国会会議録」のように長い発話を記録したのもあれば、「Yahoo! 知恵袋」のように短いものもあり、統計処理上は一つ目の課題と同じ問題が起こる。以上の問題を解決するためには、同じ基準ですべてのデータを比較できるように、データの長さをそろえる操作が必要と判断した。そこで、本研究ではすべてのデータに形態素解析を行った上、1ファイルが「100文」になるようにコンピューターソフトウェアを使って機械的に区切った。また、BCCWJの場合、4000万語の全データを解析することは計算量として膨大になることを勘案し、「書籍」と「国会会議録」では60個、「Yahoo! 知恵袋」では、全体のデータが小さいことを考慮し、30個のファイルをランダムに選択し、調査を行った。実験に使用した全データの規模を表1に示す。

学習データと評価データのいずれにも、形態素解析に基づく語彙抽出、漢字抽出を行った。その結果に基づいて、テキスト情報量や級別語彙、級別漢字の頻度などを計算し、表形式で解析の元となるデータを作成した。具体的には、以下の1から3に示す頻度情報をファイル単位で調査した。

- ◆ **変数セット1** (テキスト情報量および語種別出現頻度)：文字数、平均一文長、漢字数、漢字率 (%)⁽¹³⁾、ひらがな数、カタカナ数、英字数、外来語数、漢語数、記号数、固有名詞数、混種語数、和語数
- ◆ **変数セット2** (日本語能力試験出題基準に基づく語彙級と漢字級の出現頻度)：級外語彙数、1級語彙数、2級語彙数、3級語彙数、4級語彙数、級外漢字数、1級漢字数、2級漢字数、3級漢字数、4級漢字数
- ◆ **変数セット3** (形態素の頻度)：感動詞数、形状詞数、形容詞数、助詞数、助動詞数、接続詞数、接頭辞数、接尾辞数、代名詞数、動詞数、副詞数、補助記号

数, 名詞数, 連体詞数

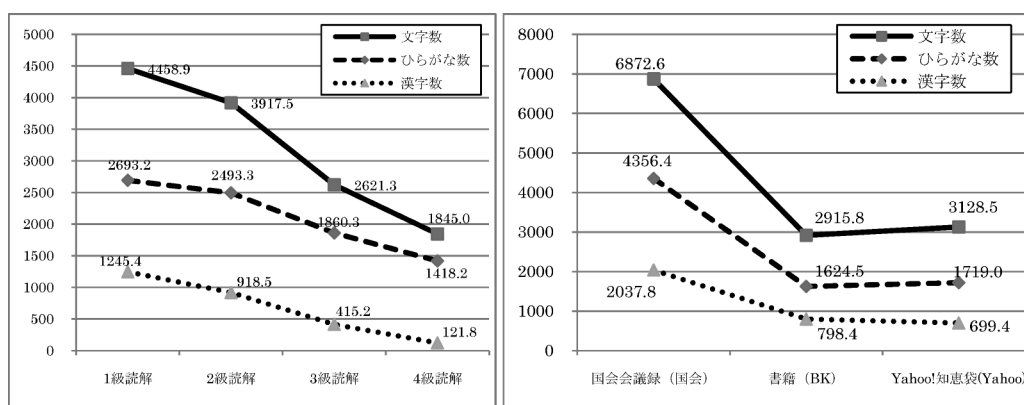
変数セット1は, 柴崎・玉岡(2010)などが行ったテキストの難易度推定の研究でよく用いられる変数であるが, 基本的な計算は形態素解析の結果に基づいて行っている。変数セット2は, 問題作成において統制している変数ということになるが, 個々のテキストに1級から4級の漢字と語彙, 出題基準外の漢字と語彙(級外漢字・級外語彙)がいくつ含まれているかを調べたものである。なお, 各級の語彙には当該級の下位級の語彙は含まれていない。そのため, 1級語彙には2級語彙が含まれておらず, 3級語彙においても4級語彙は含まれていない。最後に, 変数セット3は, 形態素の情報とテキスト特徴の関係を検討するため, 導入した変数であるが, UniDic⁽¹⁴⁾の品詞体系に準拠している。

4. 結果

まず, 前節で示した三つの変数セットに対する頻度調査の結果を報告する。紙幅の都合上, 変数セット1から3の合計37変数すべてを示すことは難しいため, ケース間の相違を明確に表すものとして, 変数セット1の集計結果を報告する。

図2では, 全データのファイル単位の平均値として, ファイル全体の文字数とひらがな数, 漢字数を示した。図2(a)によれば, 1級読解テキストの場合, 一つのファイルに平均して4458.9文字が含まれており, その中の1245.4文字が漢字であることが分かる。一方の4級読解テキストの場合, 一つのファイルの平均文字数が1845文字あり, その中の121.8文字が漢字であることが分かる。割合として見た場合, 1級の読解テキストの場合, 全体の27%が漢字であるのに対して, 4級の読解テキストの場合, 全体の僅か7%程度が漢字であるということが明らかになった。同じ観点で(b)の評価データに注目した場合, 国会会議録が特異なデータであることが示唆される。

次に, 37個の変数をもとに判別分析を行った。判別分析は「ステップワイズ法(変数増減法)」で行い, ケースの判別に寄与度の大きい変数を選び出した。3節で紹介した【第一段階】の結果として, 表2と図3が得られた。



(a) 学習データにおける文字情報量

(b) 評価データにおける文字情報量

図2 文字情報量の集計

表2では、判別に使用された関数の詳細が示されている。そして、各関数における値の大小を確認することで、ケースの判別に寄与度の高い変数が何であったのかを知ることができるが、文字数、漢字数、漢字率が大きな値を示していることが分かる。そして、大きな値ではないが、固有名詞の頻度や1級語彙の頻度、4級語彙の頻度が判別に貢献していることが分かる。次に、表2の関数1と関数2によるケースの散布図とグループの重心を確認したところ、図3が確認された。

図3では、左端の4級のテキストを起点とし、3級、2級・1級の順で重心が置かれ、その周辺にケース（読解問題テキスト）が分布している状況を確認できる。図3から読み取れることとして、（関数1によって）4級、3級、2級は十分に離れた距離に重心が位置づけられているのに対して、2級と1級は相対的に近くに布置されていることが分かる。このことは、2級と1級はテキストの特徴として多分に類似していることを示している。次に、誤判別率を確認した。「1つ取って置き法」⁽¹⁵⁾による評価を行った結果、97.7%の正答率で（観測データに対して予測データが）正しく分類されていることが明らかになり、結論的には、良い判別基準が得られたことになる。

さて、以上の結果を踏まえ、3節で紹介した【第二段階】の評価データに対する分類を行った。その結果、表3が得られた。

表3では、学習データをもとに作成した判別基準から評価データとなるBCCWJを分類した結果が示されている。これによれば、「書籍」データは、1級から3級に渡って、幅広く分布しているのに対して、国会会議録はすべてのデータが1級相当のテキストとし

表2 標準化された正準判別関数係数表

	関数		
	1	2	3
1級語彙頻度	-.164	-.537	1.104
4級語彙頻度	-.657	.125	1.451
文字数	4.625	2.563	-1.028
漢字数	-4.881	-3.560	-.810
漢字率	3.700	1.767	.338
固有名詞	-1.123	-.038	.660

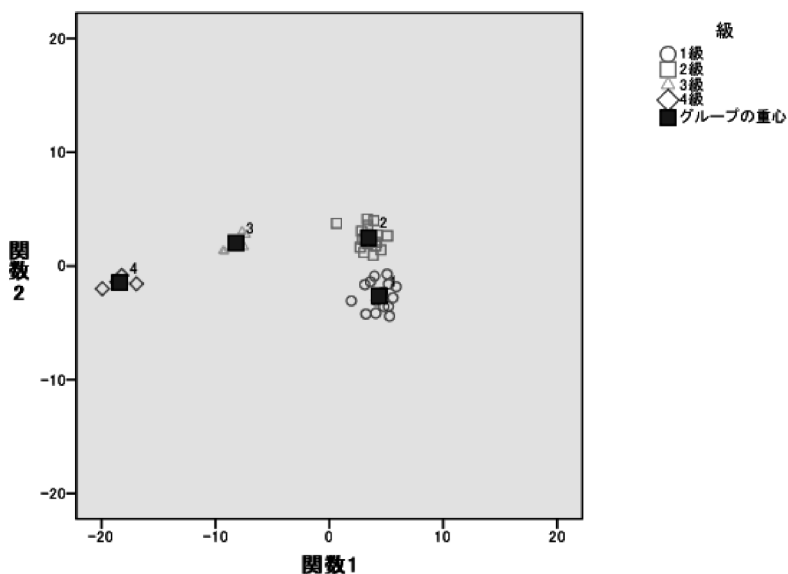


図3 全グループの散布図

表3 BCCWJ に対する分類結果

		予測グループ			合計
		1 級	2 級	3 級	
書籍	度数	17	30	13	60
	%	28.3%	50.0%	21.7%	100.0%
国会会議録	度数	60	0	0	60
	%	100.0%	.0%	.0%	100.0%
Yahoo! 知恵袋	度数	0	20	10	30
	%	.0%	66.7%	33.3%	100.0%
合計	度数	77	50	23	150
	%	51.3%	33.3%	15.3%	100.0%

て分類されている。そして、「Yahoo! 知恵袋」の場合、7対3の割合で、2級と3級に分類されている。そして、4級相当のテキストは存在しないことが明らかになった。

5. 考察

判別分析の結果を踏まえ、二点の考察を行う。一点目に、表2の結果を踏まえ、読解テキストのレベルの差を特徴づける変数は何かという問題、二点目に、表3の結果を踏まえ、「書籍」「国会会議録」「Yahoo! 知恵袋」のデータをどのように位置づけるべきかという問題について考察する。

まず、最初の問題として、1級から4級の読解テキストのレベルの差は何によって決定されているのだろうか。表2の結果から、いずれの関数においても、文字数と漢字数が全体のテキストの分類に大きく寄与していることが分かる。これに関連する興味深い事実として、問題作成者にとって、強く意識され、統制されている語彙級や漢字級の頻度（変数セット2）よりも、相対的に緩やかにしか統制されていない文字数や漢字数といった要素（変数セット1）のほうが級の判別にとって高い寄与率を示していることである⁽¹⁶⁾。この問題を追及すべく、1級語彙と4級語彙の出現頻度、文字数と漢字数の（出現頻度の実数に基づいて）散布図を作成してみた。

まず、図4の(a)に注目した場合、級が上がるにつれて、漢字数と文字数がともに増えていく様子が確認できる。しかし、(c)の語彙頻度に注目した場合、4級語彙の頻度に関しては、4級のテキストにおいてはある程度のみを形成しているものの、1級～3級においては大きな差がない。また、1級語彙の頻度に関しても1級と2級、3級の分離にとっては有効であることが分かるが、3級と4級においてはほとんど0に近い値しか示しておらず、級を判別する変数としては不十分であることが分かる。こうした分布に関連する事実として、4級語彙は基本語として位置づけられる語彙が多いため、文字数の大小に必然的に左右される。すなわち、絶対量として一文が長いテキストにおいては、4級語彙は出現頻度が高く、短いテキストにおいては、4級語彙の出現頻度が低い。このことは、(d)の評価データを見ても確認できる。つまり、「国会会議録」は(b)の通り、絶対的な文字数としても長いいため、(1級語彙のように難しい語彙も多数使用されているが)4級語彙の出現頻度が「書籍」や「Yahoo! 知恵袋」に比べ、高いことが分かる。さらに(d)に

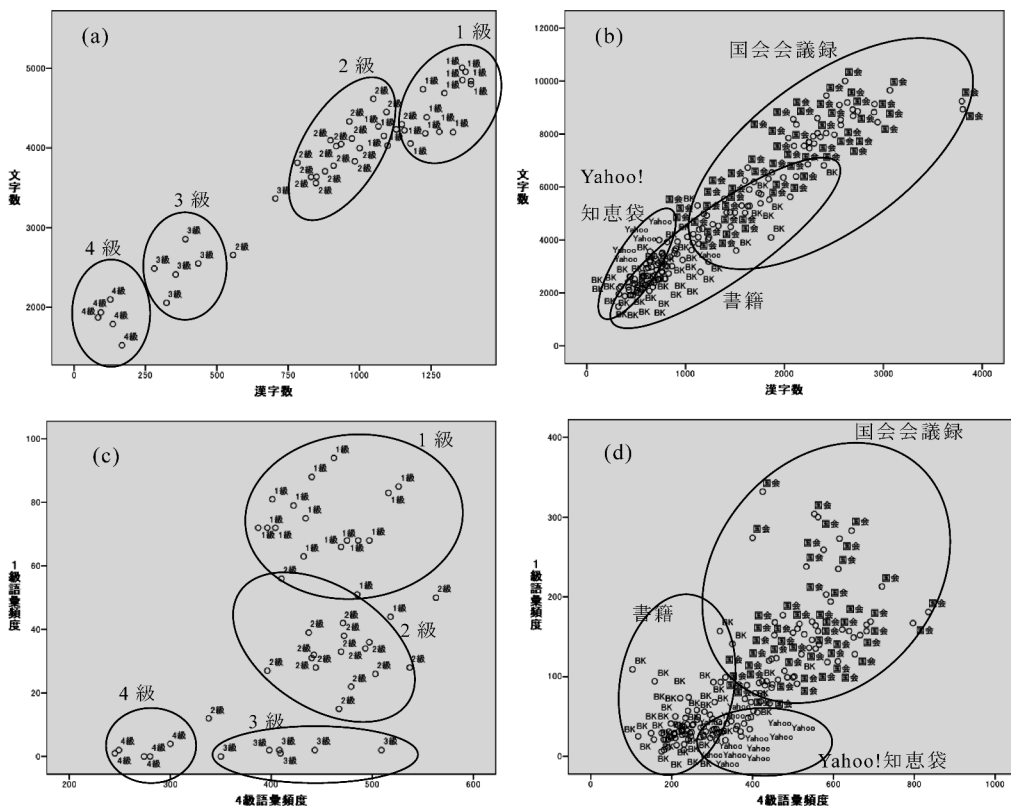


図4 頻度に基づく散布図

において、「書籍」に注目した場合、全体的に縦方向へ分布していることが確認できる一方で、「Yahoo! 知恵袋」の場合、横方向に分布していることが確認できる。すなわち、「書籍」のほうが「Yahoo! 知恵袋」に比べ、1級語彙のような難しい語彙が高頻度で使用されていることを示している。

次の課題として、コーパスデータの日本語能力試験に基づく級区分をどのように捉えるべきかという問題について考えてみたい。表3の注目すべき部分としては、「書籍」は1級から3級まで広く分布しているのに対して、「国会会議録」はすべて1級相当であるということである。このことの意味をより詳細に調べるべく、全データに対して、主成分分析を行った。その結果、6つの主成分が得られた。そこで、第一主成分と第二主成分の主成分得点を計算し、ケースの散布図を作成してみた。その結果を図5に示す。

図5では、学習データと評価データをともに布置しているが、学習データに注目した場合、第一主成分の得点において4級から1級が連続的に分布している状況が確認できる。そして、学習データの周辺に分布している評価データを見た場合、以下の3点が確認できる。

1. 「Yahoo!知恵袋」は2級、3級テキストの周辺に多く分布している。
2. 「書籍」は1級から3級テキストの全体において広く分布している。
3. 「国会会議録」はほとんどのデータが1級テキストの上に広く分布している。

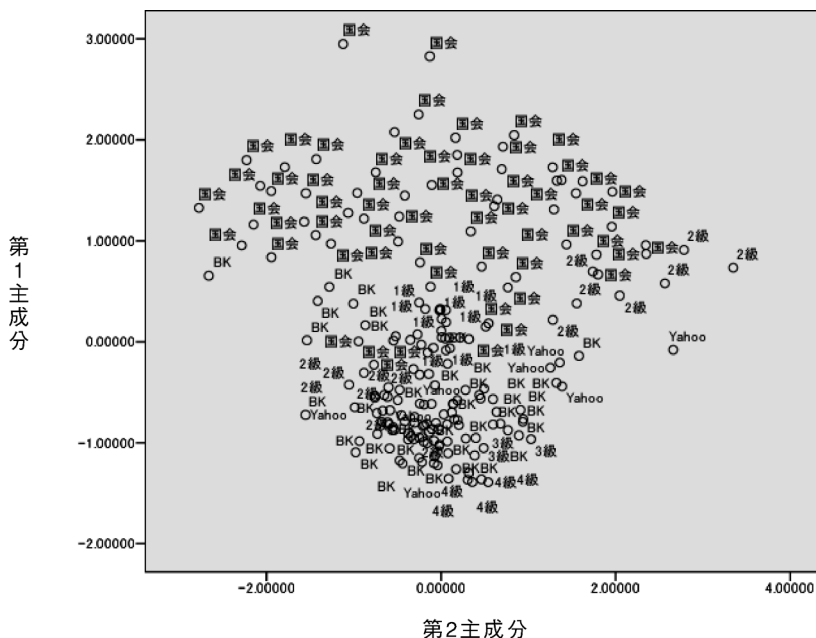


図5 主成分得点によるケースの分布

まず、1と2は判別分析の結果とも共通するもので、判別分析の結果が別の分析においても確認されたことを意味し、前節の分析結果の安定性を示している。次に3は主成分分析によって明示化できたことで、国会会議録が持つ特殊性を示している。これらの事実が示唆することとして、次の二点が挙げられる。一点目として、書籍データは多種多様なものがあるため、テキストとしての級区分ということにおいても多様なものが含まれていること、二点目として、国会会議録は1級相当のテキストというよりは1級を超えた級区分のテキストと考えるべきである。

6. まとめと課題

本稿では、日本語テストでのコーパス利用をテーマに現状と課題について紹介した。とりわけ課題の部分として、コーパスデータに対する定量的な評価がなされていないことを問題提起し、3節以下において、コーパスデータの定量的評価を試みた。その結果、日本語テストにおいて参考資料としてコーパスを活用する際のポイントとして、以下の三点が明らかになった。

- A) 「書籍」データは、日本語能力試験の級区分から見た場合、多種多様なものがあるため、個別のテキスト単位で適用可能な級区分に対する慎重な検討が必要。
- B) 「Yahoo!知恵袋」などのウェブデータの場合、2級と3級の問題検討において有効に活用できる。
- C) 「国会会議録」は、日本語能力試験のレベル感からは全体的に難しいテキストが多いため、その使用は慎重であるべき。

以上の3点に加え、テキストのレベルの差を特徴づける要因として、文字数と漢字数が相対的に重要な変数であることも明らかになった⁽¹⁷⁾。

最後に、今後の課題として二点を挙げたい。一点目に、本稿はこれまでの日本語研究で多用されたデータを日本語教育的観点から検討することが目的だったため、「国会会議録」や「Yahoo! 知恵袋」のようなある意味で特殊な書き言葉のデータも排除することなく、使用したが、日本語学習者にとって必要な言語運用という点で考えるなら、より一般的なデータに対するきめ細かい調査も必要である。例えば、本稿の調査においても、「書籍」データは非常に多様であることが確認されたので、次の調査として書籍のジャンル別にデータを選別し、本稿が行ったのと同様の手順で調査を行う必要がある。二点目に、頻度や数値に反映されない情報として話題や場面といったものを考慮した精密な観察記述が必要である。特に課題遂行のためのコミュニケーション能力の育成という観点から見た場合、個々の表現の真正性は最終的には具体化された場面に対応させながら、検討していかなければならない。この課題については、質的観察によるアプローチも含めて今後検討していきたい。また、二点目の課題は、日本語教育にとって必要なコーパスのあり方にも関連してくる問題で、将来的には話題や場面情報、さらには発話行為の情報がタグとして付与されたコーパスが開発されれば、言語教育分野におけるコーパス分析の新たな可能性が開けるであろう。

注

- (1) 多くのコーパスデータがテキストファイル (txt ファイル; Windows のメモ帳で開くファイル形式) になっているのは、この柔軟性の要件を満たすためであるが、近年は XML ファイルという新しい形式も提案されている。テキストファイルにせよ、XML ファイルにせよ、いずれもユーザーのマシン環境を選ばない点、特定のアプリケーションに依存しない点において、柔軟性と汎用性を併せ持った形式であると言える。
- (2) 英語教育の例としては、石川 (2008) において具体的な研究史および方法論などが示されている。
- (3) 90年代以降に作られた代表的な学習者コーパスとして、「International Corpus of Learner English」(国際学習者英語コーパス、略称 ICLE) や「Longman Learners' Corpus」(ロングマン学習者コーパス、略称 LLC)、中国人英語学習者コーパスである「HKUST Corpus」などがある。
- (4) 1980年に John Sinclair を編集主幹として始まった「COBUILD Project」は、コーパスと辞書編纂の密接な結び付きをもっとも鮮明に示したプロジェクトであるが、このプロジェクトによって、「Collins Cobuild Advanced Learner's English Dictionary」, 「Collins COBUILD Concise Learner's Dictionary」, 「Collins COBUILD Student's Dictionary」, 「Collins COBUILD Active English Dictionary」など、複数の辞書が公開されている。
- (5) ETS (Educational Testing Service) はアメリカにある世界最大規模の非営利テスト研究機関である。詳細は、<http://www.ets.org/> を参照してほしい。
- (6) Cambridge ESOL (English for Speakers of Other Languages) はイギリスにあるヨーロッパ最大の言語評価の機関である。詳細は、<http://www.cambridgeesol.org/> を参照してほしい。

- (7) 「Cambridge Learner Corpus」の詳細は http://cambridge.org/elt/corpus/learner_corpus.htm 参照してほしい。
- (8) バックマン, L. F., パーマー, A. S. (2000; 25) はテスト課題と得点解釈, 一般化の領域と真正性の関係を図1のように規定している。

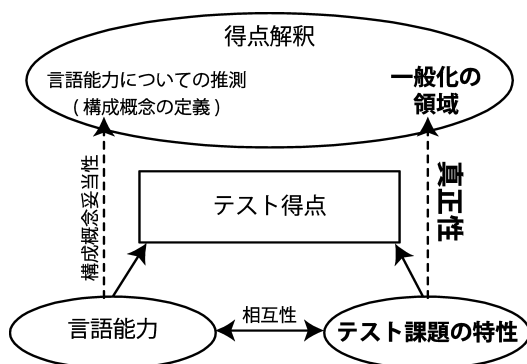


図1 得点解釈とテスト課題の特性について

- (9) <http://reed.kuee.kyoto-u.ac.jp/cf-search/> からアクセスできる。
- (10) 詳細は <http://www.tufs.ac.jp/ts/personal/usamiken/corpora.htm> を参照してほしい。
- (11) BCCWJの詳細は <http://www.tokuteicorpus.jp/> から見る事ができ、ウェブ上のデモ版は <http://www.kotonoha.gr.jp/demo/>, 構築途中のモニター版 (DVD版) は http://www.ninjal.ac.jp/kotonoha/ex_8.html から入手することができる。
- (12) 読解テキストの数量的基準に関しては『日本語能力試験出題基準 (改訂版)』(227ページ)を参照してほしい。
- (13) 漢字率とは一文中に何割の漢字が含まれているかを比率で表す。例えば「太郎は日本人だ」の場合、全体の文字は7文字であるが、漢字は5文字であるため、漢字率は「 $5/7 \times 100 = 71.4\%$ 」となる。
- (14) UniDicは機械的な方法で日本語テキストを国立国語研究所で規定した「短単位」に準拠して分割し、形態素情報などを付与する電子辞書である。なお、UniDicを使用するには、解析エンジンを別途インストールする必要がある。詳細は <http://www.tokuteicorpus.jp/dist/> を参照してほしい。
- (15) 元のデータから1つだけ外して判別基準を作り、外したデータを新たなデータとして適用した際に妥当な結果が得られるかを検証する手法である。なお、本研究で使用した統計ソフトのSPSSの場合、「交差確認済み」の分類結果を見ることで確認できる。
- (16) 読解テキストの文字数に関する規定は『日本語能力試験出題基準 (改訂版)』(228ページ)を参照してほしい。
- (17) 柴崎・玉岡 (2010) などでは一文における平均述語数などを入れた文章難易度の計算式を提案している。これは、「構文の複雑さ」と文章難易度の関連性を示す研究として位置付けられる。これに関連するものとして「語彙としての難しさの問題」、さらには、数量化は難しいと思うが「話題としての難しさ」、「文章の結束性の問題」も当然ながら文章の難

しさに強く影響するものであると考えられる。こうしたことから考えてみた場合、どこまでの要因が文章難易度に影響を与え、どこまでが影響を与えないのかの線引きは現実問題として非常に難しく、最終的な結論を出すにはもう少し大量のデータと経験事実を積み上げていかなければならない。

参考文献

- (1) 斎藤俊雄, 中村純作, 赤野一郎(編) (2005) 『英語コーパス言語学 基礎と実践』 研究社。
- (2) 石川慎一郎 (2008) 『英語コーパスと言語教育』 大修館書店。
- (3) 国際交流基金・日本国際教育支援協会 (2006) 『日本語能力試験出題基準 (改訂版)』 凡人社。
- (4) 柴崎秀子, 玉岡賀津雄 (2010) 「国語科教科書を基にした小・中学校の文章難易度学年判定式の構築」『日本教育工学会論文誌』 33 卷 4 号, pp. 449-458.
- (5) 柴崎秀子 (2010) 「日本語能力試験を土台にした文章の難易尺度の構築」『2010 年度日本語教育学会春季大会予稿集』, pp. 211-215.
- (6) 根岸雅史 (2007) 『コミュニカティブ・テストングへの挑戦』 三省堂。
- (7) 小池生夫 (他編) (2003) 『応用言語学辞典』 研究社。
- (8) 言語処理学会 (2009) 『言語処理学辞典』 共立出版。
- (9) 松田謙次郎 (2008) 『国会会議録を使った日本語研究』 ひつじ書房。
- (10) 李 在鎬, 横森大輔, 土屋智行 (2009) 「コーパス調査による形容詞の連体形と連用形」, 『コーパスを利用した国語辞典編集法の研究』 (特定領域研究日本語コーパス平成 20 年度研究成果報告書), pp. 103-110.
- (11) Alderson, J. C. (1996) Do corpora have a role in language assessment? In J. Thomas and M. Short (Eds.). *Using Corpora for Language Research*. pp. 248-259. Longman.
- (12) Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press. (バックマン, L. F., パーマー, A. S. (2000) 『〈実践〉言語テスト作成法』 (大友賢二, ランドルフ・スラッシャー監訳) 大修館書店)
- (13) Barker, Fiona. (2004) Using corpora in Language Testing. *Modern English Teacher*, 13-2, pp. 63-67.
- (14) Barker, Fiona. (2006) Corpora and language assessment: trends and prospects, *Cambridge ESOL Research Notes* 26, pp. 1-4. (<http://www.cambridgeesol.org/>)
- (15) Biber, D., S. Conrad & R. Reppen. (1998) *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- (16) Biber D., Conrad, S., & Cortes, V. (2004). If you look at. . . : Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 2004, pp. 371-405.
- (17) Chapelle, Carol A. (2001) *Computer Applications in Second Language Acquisition*, Cambridge University Press.
- (18) Chodorow, Martin & Gamon, Michael and Tetreault, Joel (2010) The utility of article and preposition error correction systems for English language learners: Feedback and assessment. *Language Testing* 27-3, pp. 419-436.

- (19) He, Lianzhen & Dai, Ying (2006) A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing* 23-3, pp. 370-401.
- (20) Hunston, Susan. (2002) *Corpora in Applied Linguistics*, Cambridge University Press
- (21) Rimmer, Wayne (2006) Measuring grammatical complexity: the Gordian knot, *Language Testing* 23-4, pp. 497-519.
- (22) Stubbs, Michael. (2002) *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

((独)国際交流基金日本語試験センター)

.....

**The Utility of Corpora for Composing Reading Comprehension
Questions for Large-Scale Tests**

LEE Jae-ho

This paper examines the utility of corpora in the composition of questions for large-scale tests. After explaining the present situation of corpus use in the field of language testing, I explain some of the advantages of using corpora in the process of composing questions. Next, in the context of reading comprehension questions, I perform a discriminant analysis of corpus data according the levels of the Japanese Language Proficiency Test, using the reading comprehension texts from the Japanese Language Proficiency Test as learning data, and data from the Balanced Corpus of Contemporary Japanese as evaluation data. Based on this analysis I identify the levels to which the Balanced Corpus data corresponds, and the extent to which such data can be used in the composition of test questions. In conclusion, I also indicate the importance of taking a qualitative approach to analysis that examines concrete examples, in addition to the quantitative approach described above.

(●●●●●)