大規模コーパスに基づく語彙リストの検証

李（り） 在鎬（じぇほ）
筑波大学

1. はじめに

近年, 国立国語研究所によって開発された「現代日本語書き言葉均衡コーパス」(Balanced Corpus of Contemporary Written Japanese: BCCWJ）を第二言語教育においても利用しようとする試みが活発に行われている。特に辞書開発における利用価値は非常に高い。見出し語の選別，語義の抽出と配置，共起語抽出と用例作成などあらゆる場面においてコーパスは，潜在的な可能性を持っていると言える[1]。このことを具体的に示す日本語教育分野の研究として，李・砂川(2012)がある。これらの研究では基盤研究（A）「汎用的日本語学習辞書開発データベース構築とその基盤形成のための研究」(代表：砂川有里子)（以下,「学習辞書科研」）で開発している「日本語教育語彙表」とそれに基づく汎用データベースの開発について紹介している。その詳細は，2 節で述べるが，特に注目すべき試みとして 1) 大規模コーパスと日本語教科書テキストデータを使った見出し語の抽出と 2) それに基づく言語情報の付与，そして 3) コーパスを利用した用例作成を行っている点である。

以下では，まず 2 節で「日本語教育語彙表」について述べたあと，3 節以下では BCCWJ の頻度情報を使い，「日本語教育語彙表」における動詞の語彙難易度に関する妥当性検証について述べる。具体的には，BCCWJ の中納言（Ver 1.0.5）を利用し，動詞の用例を収集し，「日本語教育語彙表」の語彙難易度とどのような関連を持つかを分散分析と主成分分析を使って調査を行った。調査の結果として，1) 語彙の難易度によって使用頻度上に統計的な有意差が存在すること，2) 語彙の難易度が上がるにつれ，使用頻度は下がる傾向にあることが明らかになった。

2. 日本語教育語彙表について

本節では「日本語教育語彙表」の開発背景や構築手順などについて述べる。

2.1. 開発背景

「学習辞書科研」では，データベース作成の第一ステップとしてどのような語彙を対象にデータベースを作成するかについて検討し，基準となる語彙表の構築を試みた。日本語教育分野における語彙表としては「旧日本語能力試験出題基準語彙表」(以下「旧試験語彙表」）が知られており，教育現場での指導基準，教材開発，語彙研究などの基礎資料として広く活用されているが，本プロジェクトでは，以下の理由から旧試験語彙表を採用せず，独自の日本語教育基本語彙表を構築した。

1. 日本人英語学習者のための語彙リスト JACET8000（大学英語教育学会基本語改定委員会(編) 2003）は学習者が遭遇しやすいサブコーパスと BNC コーパス頻度を対数尤度比で比較し，作成されたものである。コーパスを第二言語教育に活用することの重要性を指摘した研究として Nation（2001）があり，それによると、英語の高頻度の 2000 基本語彙が 70%～80%のテキストの内容をカバーされていることを明らかにしている。

1. 「旧試験語彙表」は作成から 30 年以上経っており，語彙の変化に対応していない。
2. 海外受験者への配慮から，文化などに関連する語彙が含まれていない。
3. 難易度設定は，テスト作成のためのものであり，教育目標ではない。

まず，1 について，「旧試験語彙表」は 80 年代に人手で作成したものであり，2 回の改訂があったものの基本的には 80 年代のものから大きく変わっておらず，新しい語彙の変化に対応していないという問題がある（cf. 押尾(他)(2007)）。具体的な問題点として，外来語が非常に少ない点や，擬音語，擬態語など表現を豊かにするための語彙項目が非常に少ない点が挙げられる。

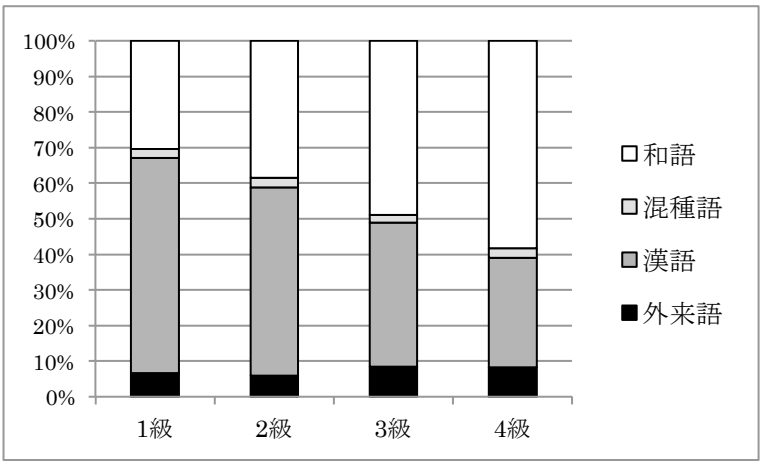

図 1. 「旧試験語彙表」における語種の分布

図 1 は，「旧試験語彙表」における語種の分布を棒グラフで示したものである。4 級語彙に関しては，和語が全体の 5 割以上を占めているが，級が上がるにつれ，漢語の比率が高くなっていることが見て取れる。語種の比率でとりわけ問題になるのは外来語で，図 1 に見られるように，いずれの級でも外来語は 1 割にも満たない低い数値である。こうした外来語の不足は，現在の日本社会における言語使用の実態とマッチしておらず，見直す必要がある。同じく，擬音語，擬態語に関しても，「旧試験語彙表」には，「にこにこ，ぴかぴか，ふらふら，わくわく」など非常に限られたものしか入っていないという問題がある。

2 に関しては，海外での受験者に対する配慮のために，「旧試験語彙表」では日本文化に関する語彙や食べ物，動植物の名前などが意図的に排除されているという問題が指摘できる。それは，日本語能力試験は言語能力を問う試験であり，文化的知識を評価の対象とするものではないという捉え方によるものである。この種のテストのためという目的に照らし合わせて考えるなら，「旧試験語彙表」の方針自体は理にかなったものであるが，日本事情や日本文化に関する教育を行なっている日本語教育の実情からはかけ離れたものであると言わざるを得ない。

最後に，3 として「旧試験語彙表」の基本的な開発用途は，あくまで学習者の日本語能力を評価するためのものであり，それ以上でもそれ以下でもないという点が挙げられる。「旧試験語彙表」はこのような前提に基づいて作成されたものであるため，それを教材開発に使ったり，辞書

開発に使うのは，本来の用途とは違うものであり，その使い方には当然ながら無理が生じる。とりわけ語彙の難易度設定（語彙級）を考えた場合，テスト問題を作成するという視点が多分に反映されており，教育目的とは必ずしも一致しないものがある。実際，テストにおける難易度設定は「知っていると想定できる日本語のレベル」という観点に基づいており，いわゆる教育目標としての難易度設定（「このレベルの人に知ってほしい日本語のレベル」）ではない。

以上に示した 3 点の問題点を踏まえ，「学習辞書科研」では，日本語教科書の語彙調査と大規模コーパスによる目標言語領域に対する定量的な調査を行い，その調査に基づいて汎用的な日本語教育用の語彙表を作成した（以下，「日本語教育語彙表」）。

「日本語教育語彙表」の具体的な狙いとしては，1）オーセンティックな語彙項目を取り入れた教育のための語彙表を作ること，2）語彙項目に対して様々な指標をつけ，辞書開発はもちろんのこと，教育現場でのニーズに答える語彙表を作ること，3）ウェブを通じて国内外のユーザーと共有できる語彙表を作ることである。以上の目的を果たすために，次の作業を行った。1）を実現すべく，コーパスデータと自然言語処理の技術を利用し，語彙調査を行った。2）を実現すべく，日本語教師による主観判定に基づいて語彙の難易度情報をつけ，「分類語彙表」に基づいて意味情報をつけることにした。3）を実現すべく，マイクロソフト社の Excel でもテキストエディタでも利用可能な電子データ（CSV 形式）で作成することにした。

2.2. 作成手順

「日本語教育語彙表」は，以下の 4 段階の作業で作成した。

1. 語彙抽出：コーパスデータを形態素解析した後，語彙を抽出した。
2. 人手による編集：ノイズを人手で除去した。
3. 主観判定：日本語教師 5 名により語彙難易度を主観的に判定した。
4. 指標作成：各語彙項目に対する意味情報やコーパス頻度情報を挿入した。

以下では，それぞれの作成方法について説明する。

2.2.1 語彙抽出

「日本語教育語彙表」開発の第一ステップとして，「日本語教科書コーパス」と「現代日本語書き言葉均衡コーパス（http://www.tokuteicorpus.jp/）」領域公開データ（2009 年度版）の中から「Yahoo!知恵袋」と「書籍」を形態素解析した後，助詞類や助動詞類を取り除き，内容語のみを抽出した。次いで，内容語の出現頻度を集計し，出現頻度 5 以上のものをリスト化した。なお，「日本語教科書コーパス」とは，筆者が研究目的で独自に構築した日本語教科書 100 冊のテキストデータであり，一般には公開されていない。この「教科書コーパス」には，国内外で使用される主要な日本語教科書が初級から上級まで均等に入っている。また，「現代日本語書き言葉均衡コーパス」とは，国立国語研究所が構築した日本語の書き言葉の均衡コーパスであるが，本プロジェクトの作業開始時点では，正式版が公開されていなかったため，領域公開データ（2009 年度版）に含まれているテキストデータを使用した。このデータの「書籍」の収録語数は，4,000 万

語であり，語彙リスト作成にとっては十分な規模である。

形態素解析においては，MeCab と UniDic を使用した。なお，語彙抽出においては，形態素解析による語彙素（短単位）に加え，形態素 N-gram[2]を使用し，複数の形態素を結合させる方法で抽出を行った。以下に具体例を示す。

(1) 2 gram の例：愛煙家、アイスコーヒー、相次ぐ、相手方、青信号
(2) 3 gram の例：一人前、網の目、幾つか、一戸建て、一度に、何時でも、いつまでも、今にも、今一つ、運転免許証、おかげさまで、お客さん、おじいさん、おじいちゃん

(1)に示した 2gram の例としては，二つの語彙素が結合してできたものである。例えば，「愛煙家」であれば，「愛煙」と「～家」，「相手方」であれば，「相手」と「方」で一語を形成しているものである。(2)に示した 3gram の例では，「一人前」であれば，「一」と「人」と「前」の 3 語彙素が一語を構成しているものである。そして，抽出後の語彙を，人手によってノイズを除去するなどの編集を行った上で，最終的に 18,010 項目の語彙リストを完成させた。

2.2.2. 主観判定

日本語教育での活用を前提にした場合，どのレベルで出すべき語彙であるかという難易度の指標が必要である。しかし，語彙の難易度指標は，機械的に決められるものではない。日本語教師の「経験」や「勘」が反映される必要がある。ところが，「経験」や「勘」といったものは主観的なものであり，必ずしも科学的な根拠に基づくものでない。そこで，本プロジェクトでは，日本語教育歴 10 年以上の教師 5 名が語彙の難易度を判定するように依頼し，その結果を統計的に処理することで，難易度に関する指標作成を行った。

主観判定者には，2.2.1.節の方法で抽出した 18,010 項目の語彙に対して 6 段階で難易度を評定するように依頼した。具体的には，初級前半，初級後半，中級前半，中級後半，上級前半，上級後半の 6 段階である。難易度評定の際には，教室学習を前提にしてどの段階で導入するかという観点から作業するよう指示を行った。

最終的には，主観判定の結果を平均値にして，6 レベルに分類した。最終的なレベルの決定においては，二つのサイクルで作業を行った。一回目には 5 名の平均値をとったあと，その平均値と各判定者の一致度を κ 値で検討した。そこで，一致度が 0.5 を下回った判定者 1 名を除外し，二回目に平均値をとり，最終的なレベルとして決定した。こうすることで，極端に違う判定をした判定者のデータを除外することができた。作業の結果として表 1 の項目が完成した。

表 1. 主観判定の結果

語彙級	項目度数	具体例
1 初級前半	426	お休み，隣，ペット，お願いします，おはようございます，私，悪い，お手洗い，お父さん
2 初級後半	800	料理，旅行，冷蔵庫，レストラン，レモン，若い，忘れる，御

[2] N-gram とは，自然言語処理の分野で提案された言語モデルで，n 個の要素による長さを持つ文字列および形態素列のことを指す。

		苦労様，いらっしゃいませ，案内
3 中級前半	2,323	生け花，意見，以降，イコール，入れ物，色んな，岩，祝う，動かす，うそつき，宇宙人
4 中級後半	6,482	医療，衣料，衣類，色紙，祝い事，違和感，インストラクター，失う，後ろ姿，訴え，角度
5 上級前半	6,401	格段，拡張，確定，格闘，合体，がっちり，かぶせる，過不足，株主，壁紙，過保護，感覚器
6 上級後半	1,578	寒天，神主，カンパ，甲板，仰天，極小，霧吹き，愚図る，くせ者，口伝え，屈伏，組曲
総計	**18,010**	

「日本語教育語彙表」と「旧試験語彙表」を対応づけた結果，表 2 の結果となった。

表 2. 「旧試験語彙表」と「日本語教育語彙表」の対応

		旧試験語彙表の級分け					総計
		1 級	2 級	3 級	4 級	未収録	
日本語教育語彙表の級分け	1 初級前半	0	4	7	375	40	426
	2 初級後半	6	79	208	341	166	800
	3 中級前半	94	921	410	105	793	2, 323
	4 中級後半	884	1, 944	93	37	3, 524	6, 482
	5 上級前半	1, 290	449	13	0	4, 649	6, 401
	6 上級後半	118	32	0	0	1, 428	1, 578
	総計	2, 392	3, 429	731	858	10, 600	18, 010

表 2 において，未収録になっているものは，「旧試験語彙表」にはないが，「日本語教育語彙表」には収録されている語彙のことであり，合計で 10,600 語ある。「日本語教育語彙表」と「旧試験語彙表」の差を考える上で，外来語の問題は重要である。というのは，「旧試験語彙表」では，1 級語彙とされている「ジャズ」「カメラマン」「ティッシュ」などの語彙は，「日本語教育語彙表」では，初級後半の語彙として分類されている。反対に，「旧試験語彙表」で 4 級語彙とされている「レコード」や「フィルム」や「ハードディスク」などは，中級後半の語彙として分類されている。この違いは「旧試験語彙表」が作成された 80 年代の使用実態との相違が反映されたためであると思われる。

2.2.3. 指標作成

「日本語教育語彙表」には，語彙項目に対する様々な指標をつけ，データベース化を行う。具体的には，以下の情報を掲載する。

1. 語彙 ID
2. 標準的な表記
3. 読み
4. 語彙難易度
5. 品詞
6. 語種

7. 旧日本語能力試験レベル
8. 意味分類
9. アクセント情報

表 3.「日本語教育語彙表」の見本

語彙ID	標準的表記	読み	語彙難易度	品詞	語種	旧試験語彙級	意味分類	アクセント情報
10	アート	アート	中級前半	名詞-普通名詞-一般	外来語		体-活動-芸術-芸術・美術	1
40	アイスコーヒー	アイスコーヒー	初級前半	名詞-普通名詞-一般 and 名詞-普通名詞-一般	外来語		体-生産物-食料-飲料・たばこ	6
109	明かり	アカリ	中級前半	名詞-普通名詞-一般	和語	2級	体-生産物-機械-灯火 | 体-自然-自然-光	0
222	足掛かり	アシガカリ	上級後半	名詞-普通名詞-一般	和語		体-関係-空間-点	3
294	厚かましい	アツカマシイ	中級後半	形容詞-一般	和語	2級	相-活動-心-自信・誇り・恥・反省	5
262	温まる	アタタマル	中級前半	動詞-一般	和語	2級	用-自然-物質-熱	4

1 は，語彙に対する固有番号である。2 は，標準的表記として『現代国語表記辞典』に準拠して作成した。3 は，標準的表記に対する読み，4 は，主観判定の結果として得られた 6 段階の語彙レベルである。5 は，UniDic の品詞体系に準拠して作成した。6 についても基本的には UniDic の出力にそって作成しており，和語，漢語，外来語，混種語，定型句の別が示されている。7 は，旧試験語彙表の語彙級，8 は，『分類語彙表』に準拠して作成した語彙の意味分類，9 は語彙のアクセント型である。

3. 調査方法とデータ

2 節の方法で開発した「日本語教育語彙表」であるが，使用したコーパスが 2009 年度のモニター版の BCCWJ であった点を踏まえ，最終版の BCCWJ との対応を検討する必要があった。この課題に対して，本研究では動詞を例に以下の方法で調査を行った。

A) データ抽出：「中納言 Ver 1.0.5（https://chunagon.ninjal.ac.jp/, 2013 年 2 月 3 日検索結果）」の品詞検索を使い，動詞の KWIC 列を収集。
B) サンブルの決定 : BCCWJ 上で頻度 200 以上の動詞を対象に日本語教育語彙表の語彙難易度情報と付与。
C) 一元配置の分散分析：語彙難易度を因子にして，サブコーパス間で（出現頻度の）平均値に統計的な有意差があるかを検討。
D) 主成分分析：サブコーパスでの出現頻度を主成分分析し，語彙難易度別に検討。

A）の方法で，全コーパスをターゲットにし，「1,122,649」用例を収集した。B）の方法で，動詞のタイプ頻度「465」，トークン頻度「495,969」の用例を抽出し，各見出し語に対して「日本語教育語彙表」における「初級前半」から「上級後半」の 6 段階の語彙難易度を抽出した。抽出の結果を表 3 に示す。

表 4. 語彙難易度×BCCWJ における頻度

語彙難易度	タイプ頻度(%)	トークン頻度(%)	一語の平均使用度数
1（初級前半）	22(4.7%)	104,171(21.0%)	4735.05
2（初級後半）	61(13.1%)	126,791(25.6%)	2078.54
3（中級前半）	121(26.0%)	95,317(19.2%)	787.74
4（中級後半）	199(42.8%)	138,136(27.9%)	694.15
5（上級前半）	56(12.0%)	29,290(5.9%)	523.04
6（上級後半）	3(0.7%)	945(0.2%)	315.00
級外	3(0.7%)	1,319(0.2%)	439.67
総計	465(100.0%)	495,969(100.0%)	1066.6

表 4 には，B）で抽出した BCCWJ 上で頻度 200 以上の動詞が「日本語教育語彙表」でどのような語彙難易度として分布するかを示している。例えば，初級前半の見出し語は，22 項目あり，その延べの出現頻度は，104,171 である。なお，級外とは，「日本語教育語彙表」における未収録語で「際する，売れる，存ずる」の 3 語である。

4. 結果と考察

3 節の C）で一元配置分散分析を行った結果，表 4 が明らかになった。

表 5. 分散分析の結果

サブコーパス	平均平方	F 値	有意確率
コア出版・雑誌	48899.398	9.749	.000
コア出版・書籍	91697.020	9.589	.000
コア出版・新聞	31928.625	7.364	.000
コア特定目的・ブログ	12672.004	19.344	.000
コア特定目的・知恵袋	28195.890	18.417	.000
コア特定目的・白書	1518.605	.317	.903
非コア出版・雑誌	261404.139	13.137	.000
非コア特定目的・広報誌	31881.073	3.560	.004
非コア特定目的・教科書	21731.399	10.701	.000
非コア特定目的・韻文	1571.395	22.206	.000
非コア出版・書籍	1021625.025	11.019	.000
非コア出版・新聞	334925.015	8.135	.000
非コア図書館・書籍	1029793.560	15.094	.000
非コア特定目的・ブログ	401148.434	31.093	.000
非コア特定目的・知恵袋	1258939.389	16.776	.000
非コア特定目的・ベストセラー	230904.189	11.582	.000
非コア特定目的・白書	2198.841	.207	.959
非コア特定目的・国会会議録	7186582.150	4.636	.000
非コア特定目的・法律	16231.052	.674	.643

表 5 から白書と法律を除くすべてのサブコーパスにおいて語彙難易度によるサブコーパスにおける出現頻度の平均値に有意な差が確認できる。平均値の推移は，図 2 のとおりである。

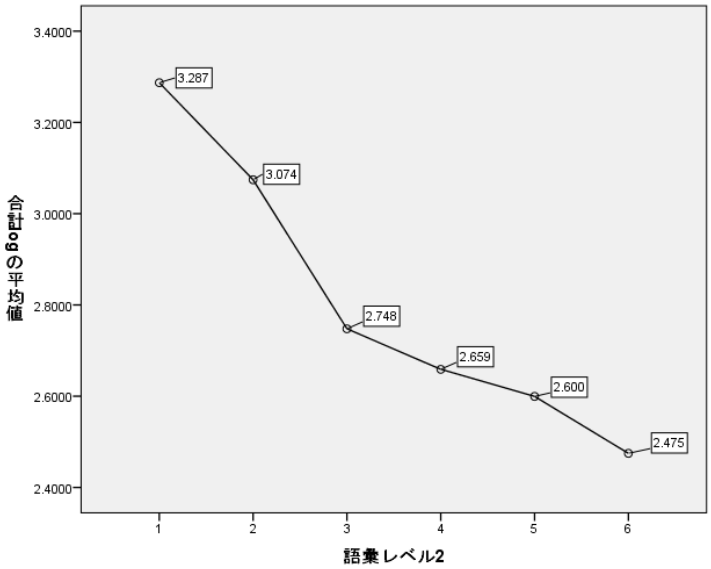

図 2. 語彙難易度による BCCWJ の合計頻度の対数変換済み値

図2は, 語彙難易度別のBCCWJ合計頻度の推移を示している。この図から確認できることとして, 語彙難易度が上がるにつれ，出現頻度の平均値が下がっている実態が見て取れる。段差を見た場合, 1(初級前半)と2(初級後半), 3(中級前半)から5(上級前半), そして6(上級後半)で差が見られる。1と2は，いわゆる基本語に相当するものと捉えることができ，高頻度語が集中していると捉えることができ，3から5は中程度の頻度，6は低頻度語として解釈できる。さらに詳細な分析をすべく，出現頻度を説明できる合成変数（主成分）の抽出を試みた。主成分分析の際には，分散分析の結果を受け，分析変数として法律と白書は削除した。結果は表6のとおりである。

表6. 抽出した主成分

	主成分	
	第一	第二
コア出版・雑誌	.868	.036
コア出版・書籍	.887	-.077
コア出版・新聞	.772	.492
コア特定目的・ブログ	.845	-.208
コア特定目的・知恵袋	.797	-.303
非コア出版・雑誌	.896	.061
非コア特定目的・広報誌	.619	.515
非コア特定目的・教科書	.759	.255
非コア特定目的・韻文	.537	-.429
非コア出版・書籍	.879	.074
非コア出版・新聞	.787	.479
非コア図書館・書籍	.816	-.072
非コア特定目的・ブログ	.878	-.239
非コア特定目的・知恵袋	.783	-.347
非コア特定目的・ベストセラー	.831	-.197

主成分分析におけるKMOの標本妥当性は「.934」とかなり高い値を示しており，分析の妥当性を

示す。得られた主成分を解釈すると，第一主成分は，コアデータの出版・書籍，非コアデータの雑誌，非コアデータのブログにおける出現頻度に高い得点を与えるものになっている。また第二主成分は，非コアデータの広報誌の出現頻度に高い得点を与えるものになっている。なお，第二主成分までで累積固有値は73.5%となっており，この二つの主成分でほとんどのデータの分布が説明できることが明らかになった。そして，主成分得点をもとに，各ケースの散布図を作成した。散布図作成時には，見やすさを考慮し，1～3の下位級と4～6の上位級に分けた。

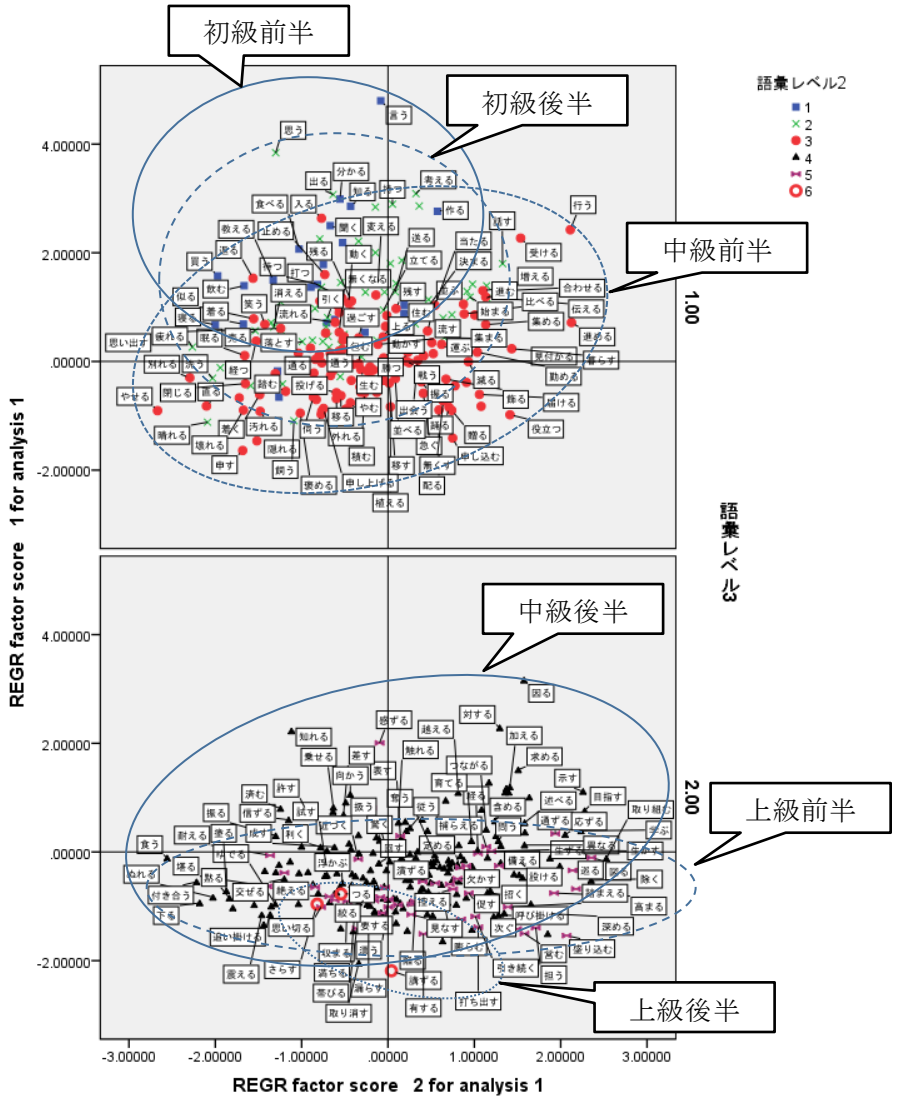

図3. 第一主成分得点×第二主成分得点の散布図

図3における参照線に基づいて解釈すると，1(初級前半)は第一主成分において高得点を示し，少ない項目がもっとも高い位置を示している。次に，2(初級後半)は，項目数が多く，1より下の位置，すなわち1より低い頻度で分布している。3と4は，もっと項目数も多く，広範囲に散らばっている。5と6はいずれも，0を下回るところに位置している。図3の結果から，難易度が上がるにつれ，第一主成分得点の値が小さくなり，下位方向への緩やかな移動が確認される。

5. まとめと課題

本研究では，「学習辞書科研」で開発している「日本語教育語彙表」について説明を行った上，大規模コーパスを使った妥当性検証の一つとして，動詞を例に頻度分析を行った。分析の結果，語彙難易度による出現頻度の有意な差が観察され，語彙難易度が上がるにつれ，頻度が低くなる傾向が確認された。 今後の課題として，動詞以外の語彙に対しても順次，調査を進め，項目の確認と妥当性検証を行う。

【参考文献】

Nation, P. (2001). Learning vocabulary in another language. Cambridge University Press「語彙の計量」『講座日本語の語彙 1 語彙原論』225-243，明治書院

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向けて」『国際交流基金日本語教育紀要』第 4 号、pp71-86、国際交流基金日本語国際センター.

砂川有里子(2012)「学習辞書編集支援データベース作成について -『学習辞書科研』プロジェクトの紹介」『日本語教育連絡会議論文集』24，164-169.

李在鎬・砂川有里子(2012)「コーパスを活用した日本語語彙表の構築」2012 年日本語教育国際研究大会（ICJLE2012）パネルセッション 日本語教育につながるコーパス研究―現状と今後の展望―（名古屋大学）