博士論文審査及び最終試験の結果（案）

審査委員（主査）　佐　野　　洋

学位申請者　　本田ゆかり

論　文　名　　大規模コーパスに基づく日本語教育語彙表の作成

【審査の結果】

（審査概要）

本研究の目的は，日本語学習のための語彙リストを作成することである。大規模日本語コーパスを用い，語彙の重要度を複数の統計指標を使って定量化し，数値でランキングされた語彙の自動収集を行った。さらに語彙の難易度を日本語教育分野の観点（定量基準）から解釈することで語彙セットの再配列を実施した。この一連の手続きから作り出された基本語彙リスト内容は，日本語教育における語彙表内容の改善につなげる狙いを持ち，テキストカバー率を用いて評価され，有効な結果を得た。コーパスの語彙調整手続きや語彙のカバー範囲にいくつか問題が認められるが，本研究が，語彙の取り扱い量を格段に拡大させることで，語彙表に質的変化をもたらすことを目指した意欲的な取り組みであって，研究成果の公開が日本語学習分野に大きなインパクトを持つことを了解し，本委員会は，全員一致で博士の学位（学術）を授与するに相応しいとの結論に達した。

なお，審査委員会は，佐野洋を主査とし，本学，根岸雅史教授，吉冨朝子教授，投野由紀夫教授，学外から山内博之教授（実践女子大学文学部）を迎え，以上の４名を副査とする５名で構成した。

【論文の概要】

本研究の目的は，日本語学習のための語彙リストを作成することである。とりわけ日本語の書き言葉を理解するための基本語彙を目指した。大規模日本語コーパスを用い，語彙の重要度を複数の統計指標を使って定量化し，数値でランキングされた語彙の自動収集を行った。さらに語彙の難易度を日本語教育分野の観点（定量基準）から解釈することで語彙セットの再配列を実施した。この一連の手続きから作り出された基本語彙リスト内容は，日本語教育における語彙表内容の改善につなげる狙いを持ち，テキストカバー率を用いて評価され，有効な結果を得た。

日本語教育分野では，従来，いわゆる語彙調査の結果を参照資料とし，当該分野に携わる専門家による判定と選択という手続きにより複数の語彙表が作られてきた。発案から作成までに手間も時間もかかる。その一方で，こうした語彙表が教育現場や研究など様々な目的で利用されることについては問題点も指摘されてきた。その後，2009年に，国立国語研究所 コーパス開発センターから『現代日本語書き言葉均衡コーパス（BCCWJ）モニター版』（約3000万語）が公開された。このデータ公開以後，日本語教育分野においてコーパスから語彙を採収することで語彙表が作られ始めた。

本研究は，2011年に公開された『現代日本語書き言葉均衡コーパス（BCCWJ）』（1億450万語）（以下，BCCWJ）を語彙の採収先とした語彙リストを作成することにある。控えめなタイトル「日本語教育彙表の作成」からは分かりづらいが，語彙表の内容（語彙セット）に質的変化を狙った野心的な研究であ

る。1億語規模の言語データを取り扱う点に加えて，語彙の選択手続きに統計量（頻度，散布度，有用度）を用いることで客観的な語彙の採択基準を提案している。そして，これらの基準に則った語彙表作成を進めることで，作成手続きの一貫性を確保し，語彙表の内容に妥当性を与えている。参照する語彙の取り扱い量を格段に拡大させることで，語彙表に質的変化をもたらすことを目指している。

本研究により作り出された語彙リストは，日本語学習・教育分野での利活用が見込まれるほか，教科書テキストやテストの作成，さらには言語能力試験の内容評価に適用し，資することが期待される。

コーパスを用いて定量的なアプローチを使って語彙表を作成する場合，コーパスに含まれるテキストジャンルのバリエーションや，それらサブコーパスの語彙量の全体コーパスサイズに対する比（バランス）が語彙の出現頻度に直接影響する。そのため，利用するコーパスが，作成対象の語彙表の目的に適うかどうかを事前に検討することは極めて重要である。これまで，利用コーパスの語彙バランスを調整するなどの方法を使い語彙表作成を実施した類例は過去に見ない。また，日本語教育分野では，基本語彙選定や語彙の難易度判定は専門家判定方式で行われるべきであるという考え方を背景に，コーパス準拠の語彙表であっても，最終的な語彙採録の判断は専門家判定方式で行ったり，過去に専門家判定方式によって作られた語彙表を頼りに参照したりするなどしてきた。

日本語教育用の語彙表を開発する新しい試みとして，本研究では，語彙表の利用目的（書き言葉の日本語を理解すること）に合わせてコーパスを再構成し，語彙採録に於いては，専門家判定方式を用いず，客観的な選定基準を設け（複数の統計量を提案し），その基準の則った手続きで語彙収集を行った。語彙表規模は1万語である。2000語単位でレベル区分した。テキストカバー率調査を通じて，本語彙表の内容（語彙セット）を評価したところ，日本語教育用のテキストに対しても，日本人向けに書かれた一般のテキストについても高いカバー率を示した。さらに，話し言葉を主とするテキストについても書き言葉テキストと同様に高いカバー率を示した。

本論文（165頁）の構成は以下である。研究成果として「日本語教育基本語彙1万語リスト －読んで理解するための重要語彙－」（206頁）が付随する。この1万語リストはエクセルで検索可能なデータとしても用意されており，日本語教育現場での実利用に直ちに供することも付言する。

第1章「はじめに」は，研究の目的と意義を述べている。コーパス（電子化された大規模な言語資料）に基づき，定量的な判断基準で極力客観的に語彙を区分けること，こうして整理された語彙群を対象として，日本語教育的観点からの修訂を施した語彙表を作り上げることの目的を確認し，当該目的を支える研究の意義について記している。

第2章「先行研究の概観」は，先行研究を概観し，本研究の意義の観点から評価している。これまでに作り出された語彙表を，日本語教育の需要などの背景も踏まえながら追っている。まず，語彙調査結果を元資料として作り出された日本語教育語彙表について解説する（これを従来型と称する）。次に，英語教育分野と日本語教育分野に於けるコーパス指向の語彙表について概観している（両言語とも。但し，日本語の構築事例が少ないことも示している）。最後に，語彙調査や日本語コーパス研究の中で，同時進行で展開される日本語の語の単位の研究や漢字表記の諸研究にも目配りしている。

日本語教育基本語彙の選定活動は，戦前にまで遡ることができる。戦後，現代日本語の語彙調査は1950年頃から実施されている。従来型の日本語教育基本語彙表は，戦後の語彙調査に基づき，教育やことばの

専門家が判定するという方式で語彙が選定され，教育的配慮からレベル分け等が為されたものであった。わけても日本語教育の現場や研究で再三使用されている語彙表が，日本語能力試験のために作られた「日本語能力試験 出題基準（国際交流基金・日本国際教育支援協会編集，2007)」である。一方で，この「出題基準」が，広く教育現場や研究目的に用いられることについては問題点も度々指摘されている。

また，従来型の手続きで作り出された日本語教育語彙表は，複数の語彙表間で語彙の一致率が高くないことが先行研究で示されている。この先行研究の成果は，専門家の判断に基づく選定にはバラツキがあること，拠って日本語学習者に適した基本語彙という概念は確定的で決定的でないことを示している。人為の判断は，その結果の相互比較において不明瞭さを露呈し，その方法論や選択プロセスに限界があることを示唆している。

英語教育分野における定量的な判断基準開発の動きを概観している。英語学分野ではコーパス開発も日本語よりも先行している。こうした経緯から，英語教育に資する語彙表も，コーパスに基づいた定量基準（出現頻度や語彙分布等）に従って客観的に基本語彙を選定する試みは早くからあった。そのため，コーパスを分析対象とする各種の統計指標についての研究も進んでいる。先行する英語教育分野の語彙表開発の知見を踏まえつつ，近年では日本語教育分野でもコーパスに準拠した語彙表の開発が試みられてきている。

第 3 章「研究方法 －コーパスに基づく日本語教育語彙表の作成方法」は，本論の研究方法を説明する。まず，作り上げようとする語彙表の総語数，適用対象者，利用の範囲などの語彙表のデザインについて示した。次に，用いるコーパスが，日本語教育語彙表の言語資源（リソース）として適切か否かを検討し，コーパス内の語彙バランス（コーパスは複数のサブコーパスから構成されることが多い）を調整する方法を説明している。デザインを適用することで，合目的に再構築したコーパスを使い，語彙の頻度リストを作成するとともに，語彙の散らばりを示す散布度，散布度を基準とした有用度の有効性を解説している。これらの指標から語彙の重要度を計算する方法について述べている。また，統計指標によって選定した語彙を日本語教育的観点（単語親密度，語彙の表記，教授する文型との関わり）から水準調整する方法と具体的なレベル区分の方法を説明している。最後に，作り出される語彙表を評価する目的で，語彙のテキストカバー率の産出方法とその調査方法を述べている。

第 4 章「コーパスに基づく日本語教育語彙表の作成」では，3 章で示した研究方法と作成手続きに従い，コーパスから日本語教育語彙表を作成した結果を示している。語彙表の作成は，コーパスの採用，コーパスの語彙構成の調整を経て，各種統計量の算出を行い，コーパスから得た語彙のランキングとレベル分けの順に行った結果を顕している。

コーパスは，国立国語研究所 コーパス開発センターが提供する『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）を用いた。規模は 1 億 430 万語である。BCCWJ は 13 分野の異なるテキストジャンルを含むサブコーパスから成る。この 13 媒体の語彙の重なりを分析したところ，どの媒体にも偏在し，しかも高い頻度で出現する語彙は多くない。分けても BCCWJ の中で「特定目的サブコーパス」（白書，国会会議録，ベストセラー，Yahoo!知恵袋，Yahoo!ブログ，検定教科書）の語彙は他分野の語彙群と重なりが少なく，書籍や新聞，雑誌などは，他分野相互で重なりが比較的多いことが判明した。さらに，各分野の特徴語や語彙分布の傾向を分析したところ，国会会議録，白書と，広報誌，ベストセラー，Yahoo!知恵袋，検定教科書には専門用語が多く，中でも，国会会議録と白書と広報誌にあっては，日本語教育の視点から判断するとさほど重要ではない語彙が散見された。この調査結果を踏まえて，これら 6 媒体の語

数を減らすことで，コーパス内の語彙バランスを調整する。この結果が BCCWJ の再構築である。コーパス規模は約 9 千 5 百万語（短単位）となった。こうして調整過程を経て用意した言語資源は，先行研究で用いられた語彙調査結果に比較すると非常に大きく，コーパスを利用した先行事例に比しても最大規模である。語彙を採録する対象としての資料サイズとして十分であると考えられる。語彙群の多様性についても，日本語教育を目的としたコーパスとして検討し調整されたデータとなっている。

次に，再構築したコーパスデータを用いて語彙の頻度集計を行い，この頻度情報をもとに前章で示した散布度，有用度などの統計量を算出した。散布度は，複数の指標を検討した結果として，DP (Gries, 2008) を用いた。有用度は，DP の逆数に頻度の対数を掛けることで得ている。加えて，日本語教育で用いる観点から，語彙の難易度を示唆するものとして単語親密度を全語彙に付与している。

次のステップでは，頻度，散布度と有用度から各語彙をランキングし，日本語教育上の視点での補正を加えた。語彙群はレベル分けされ（2000 語ずつの区切りで，2000 語レベルから 1 万語レベル），有用度指標順にランキングされている（単語親密度でランク調整）。ただし 2000 語レベルの基本語彙に関しては，頻度ランク 100 位までの高頻度語彙と初級文型と関わりの深い語彙については，その単語親密度に関わらず 2000 語レベルに配置している。

第 5 章「語彙表の評価」では，作り出した語彙表を評価するため，語彙のテキストカバー率調査を実施し，その結果を示す。テキストカバー率調査では，日本語教育分野で用いることが妥当と思われるテキストとして日本語能力試験読解過去問題と，中・上級者用に書き下ろされた小説・エッセイを用いた。さらに日本語学習者が読む可能性があるテキストとして，日本人向けに書かれた一般文書を加えた。それら文書は，新聞，小説，Web サイトのテキストと話し言葉コーパスである。評価では，日本語能力試験の旧「出題基準」の語彙表のカバー率と比較を行った。その結果，凡そ，すべてのテキストにおいて，本研究で作成した語彙表の語彙セットが，より高いテキストカバー率を示した。

本研究で作成した語彙表の語彙の日本語能力試験 1 級読解過去問題におけるテキストカバー率は 93%，中・上級読解教材では 91%，新聞や小説などの一般のテキストでは 85%以上であった。この結果は，本研究で選定した 1 万語が，上記のような書き言葉テキストを理解するための日本語教育基本語彙として十分に機能することを示している。また，本研究では書き言葉コーパスをベースに語彙表を作成したが，話し言葉の語彙も同様にカバーする結果となった。

第 6 章「本研究の結論と展望」では，本研究の位置づけと研究成果を改めて確認している。本研究が示した語彙表作成における語彙の定量評価基準や選別の手続きは，先行研究において類例がない。本研究の語彙表の利用可能性としては，教材作成における利用，テスト（テスト問題作成や，作文の評価等での）利用，そして得られた語彙群そのものの研究利用がある。今後の課題として，複合語の抽出，日本語教育分野で有用な表記の併記，日本語教育的観点からのデータ補正の検討がある。さらに，クラスター分析で一つの閾値となった 6000 語（中頻度）水準以降の語彙レベルの検討，他のコーパス準拠の語彙表との一致率調査が残った。

なお，研究成果である「日本語教育基本語彙 1 万語リスト」には，各語彙の重要度（ランク）や，2000 語レベルといった語彙のレベル情報に加えて，和語/漢語の区別，単語親密度や品詞分類なども掲載されており，日本語教育現場での利用に直ぐさま役立つよう工夫されている。

【講評】

本論文の優れた点を挙げる。

・大規模なコーパスから，定量化された基準に基づいて語彙を選別する手続きを経て語彙表を作りだしたことは高く評価できる。1 億語規模のコーパスを用いたことにより，語彙の取り扱い方が量的に変化し，その結果，語彙表の作成結果に質的変化をもたらしたと考えられる。
・本研究の語彙表の語彙セットを使って計測したテキストカバー率が，日本語能力試験の旧「出題基準」の語彙表（国際交流基金・日本国際教育支援協会編集，2007）の語彙セットでのそれを超えたことを積極的に評価したい。日本語能力試験は，毎回約 60 万人の受験者をコンスタントに集める，日本語教育における最大規模の試験である。現在の出題基準は公開されていないため，一般には旧「出題基準」の語彙表しか目にすることができないのであるが，カバー率において，その語彙表を上回ったことは，特筆に値すると言ってよい。
・従来の語彙表の特徴（サンプル語彙の対象が限定されていたり，専門家の内的基準（判断や判定）による選択であったり，教科書語彙を対象にしたテスト対応であったりしたこと等）の問題を，7 つの代表的な語彙表間における語彙一致率の低さを実証的に明らかにしたこと。
・上記の課題点を鑑みた上で，公開されている大規模コーパスを使い，語彙選択の基準（幾つかの統計量）を提示し，その基準に則って語彙を選択したことで，再現性のある語彙表作成の手続きを示すことができた。属人性を排する語彙表作成の手続きを示すことができた。
・サブコーパスの語彙分布調査（クラスター分析）の結果から 6000 語程度に一つの水準があることを示した。6000 語水準を超えると，サブコーパスが違うと語彙のレベル間移動が激しくなることを実証的に示した。

一方，指摘された問題点や不備については以下である。

・説明文章の中で，例えば，どの言語（英語の事例を参照するのか，日本語のことを述べているかなど）のことについて述べているのか判断に迷う点がある。「外国語」と漠然と記載している個所があるが，表現として曖昧である。これら論文の外形的な側面として，文章の表現上改善する点があり，見直しを期したい。
・分布度基準に基づいて判断した共通性が低いサブコーパス群の語彙を減らしている。判断に使う語彙分布度と語彙調整の考えはよいが，具体的な増減手段（平均からのズレの定数倍から語彙数を決める方法）に問題がある。
・語彙の単位が短単位（語彙素）である。複合語など基本語彙としてよい語彙が欠落する結果になっている。同じ議論だが，機能語と実質語の区別を明確にしてテキストカバー率を評価するべきだろう。
・固有名詞の排除の仕方から見ると，語彙は個人の利用・運用の背景が多いだろうから，単に固有名詞の排除だけでなく，語彙の単位のことも含めて，一般的な判断基準を提案するべきではないか。
・漢字圏の学習者か，非漢字圏の学習者かによって語彙表も変化すると考えられるが，本研究の語彙表はその点が考慮されていない。今後の課題として実現を望む。
・本論の成果（語彙表の作成）を踏まえて，今後，どのように発展・展開させてゆくのかの記述が少ない

のではないか。

2016年2月21日（日）に行われた最終試験では，本田ゆかり氏からの概要説明に続き，以上に挙げた点を含む疑問点や問題点を中心に質疑応答が行われた。本田氏は，各質問に対して回答内容の根拠を示しながら丁寧に応答した。幾つかの問題点に対しては，指摘内容への同意とその理解の上で，今後の研究課題に結び付ける積極的な姿勢が示された。

【総合評価】

学位請求論文の内容，最終試験における応答を鑑み，総合的に判断した結果，本審査委員会は，全員一致で，本論文が，博士（学術）の学位を授与するに相応しいとの結論に達した。

－以上－