論文の和文要旨

| | |
|---|---|
| 論文題目 | 大規模コーパスに基づく日本語教育語彙表の作成 |
| 氏名 | 本田ゆかり |

本研究の目的は，大規模日本語コーパスを用いて，語彙の重要度を統計指標によって定量化し，難易度を日本語教育の観点で解釈して，日本語学習のための語彙リストを作成すること，その成果を用いて日本語教育語彙表の改善につなげ，日本語の書き言葉を理解するための基本語彙を選定することにある。作成した語彙表の評価としては，テキストカバー率調査を行った。

日本語教育では，従来，語彙調査の結果を資料とし専門家の主観的選定によって複数の語彙表が作られてきた。しかし，そのような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いた。その後，2009 年に国立国語研究所より『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）モニター版が公開された。それ以降，日本語教育でもコーパス準拠の語彙表が作られ始めた。

コーパス準拠で語彙表を作成する場合，コーパスに含まれる媒体のバリエーションやサブコーパスのバランスが語彙の出現頻度に直接影響する。そのため，コーパスが語彙表の目的に合っているかどうかを検討することは非常に重要である。しかし，コーパス自体の検討を十分に行ったうえで，必要に応じてコーパス自体のバランスを調整するなどして語彙表作成に利用した例は過去にない。また，日本語教育では，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参照するなどしている。

そこで本研究では，日本語教育語彙表開発の新しい試みとして，目的に合わせてバランスを調整したコーパスに基づき，専門家判定方式を用いず，客観的選定を主軸とした手法で，主に書き言葉の日本語を理解するための日本語教育基本語彙を 1 万語選び，レベル分けして提示した。そして，この語彙表を評価するためにテキストカバー率調査を行ったところ，日本語教育用のテキストでも，日本人向けに書かれた一般のテキストでも，本研究で選定した基本語彙は高いカバー率を示した。さらに，本研究の語彙表は日本語の書き言葉を理解することを主な目的としたものだが，話し言葉においても書き言葉同様に高いカバー率を示した。以下，本研究各章の概要を記す。

第 1 章では，本研究の目的と意義について述べた。本研究の目的がコーパスに基づい

て客観的に語彙を選定し日本語教育的観点からの補正を加えた日本語教育語彙表を作成することであることを確認し，本研究の意義について記した。

第 2 章では，先行研究を概観した。先行研究では，過去に行われた語彙調査を資料として作られた従来型の日本語教育語彙表について述べた。次に，英語教育と日本語教育におけるコーパスに基づく語彙表について概観した。最後に，語彙調査や日本語コーパス研究の中で進められた日本語の語の単位の研究や，漢字表記の研究について示した。

日本語教育基本語彙の選定は古くは戦前から行われていた。また，現代日本語の語彙調査は 1950 年ごろから行われ，従来型の日本語教育語彙表は，このような語彙調査に基づいて専門家の判定方式で選定され，レベル分けなどがされたものであった。その中でも，日本語教育の現場や研究でよく使われているのが「出題基準」である。しかし，日本語能力試験のために作られた「出題基準」が，広く教育現場や研究目的に使用されることについては問題点も指摘されている。また，このようにして作成された複数の日本語教育語彙表間の語彙の一致率は高くないことも先行研究では示されている。このことは，日本語学習者にとって何が基本語彙かという概念は曖昧なものであり，専門家による主観的選定には限界があることを示唆している。

一方，英語教育においては，コーパスに基づきその出現頻度や語彙分布などをもとに客観的に基本語彙を選定しようとする動きが早くからあった。そのため，コーパスにおける語彙分布を示す統計指標についても研究が進んでいる。そのような英語教育における教育語彙表開発の知見に基づき，近年は日本語教育でもコーパス準拠の語彙表が開発されるようになった。

第 3 章では，研究方法について説明した。ここではまず，語彙表の総語数，対象者，利用範囲などの語彙表のデザインについて示した。次に，コーパスが日本語教育語彙表の元データとして適切かどうかを検討し，コーパスバランスを調整する方法について説明した。そして，そのようにして再構築したコーパスを頻度リスト化し，散布度，有用度を使って語彙の重要度を定量化する方法について，また，統計指標によって選定した語彙を単語親密度，語彙の表記，文型との関わりという日本語教育的観点からランク調整し，レベル分けを行う方法について述べた。さらに，語彙表の評価として，語彙表の語彙のテキストカバー率を調査する方法について述べた。

第 4 章では，コーパスに基づく日本語教育語彙表を作成し，その結果を示した。語彙表の作成は，コーパスの選定と再構築，分析用基礎統計の算出，語彙のランキングとレベル分けの順に行った。

まず，コーパスの選定と再構築を行った。コーパスは BCCWJ を使用した。BCCWJ は 13 媒体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，どの媒体にも安定して高頻度で出現する語彙はそれほど多くないことがわかった。また，特に，BCCWJ の中でも「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）の語彙は他媒体

との重なりが少なく，書籍や新聞，雑誌などは，他媒体との重なりが比較的多いことがわかった。さらに，各媒体の特徴語や，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いことが明らかになった。そして，中でも国会会議録と白書と広報誌では特に日本語教育的にはあまり重要ではない専門用語が多いことが示された。この結果を踏まえ，これら 6 媒体の語数を削減することによってコーパスバランスを調整し，日本語教育語彙表作成のための元データとなるコーパスとして BCCWJ を再構築した。その結果，コーパスの規模は約 9 千 5 百万語（短単位）となった。このように調整したコーパスは先行研究で作られた既存の語彙表に利用されたものと比較しても最大規模のものであり，コーパスサイズとしては十分であるものと考えられる。また，内容的にも日本語教育を目的としたコーパスとして検討，調整されたものとなった。

そして，このようにして再構築したコーパスを頻度集計し，この頻度情報をもとに散布度，有用度などの分析用基礎統計を算出して付与した。散布度は複数の指標を検討した結果，DP（Gries, 2008）を使用した。有用度は DP に基づく独自の指標を使い，DP の逆数に頻度の対数を掛けて算出した。さらに，日本語教育的観点からの語彙の難易度を示唆するものとして単語親密度を付与した。

次に，頻度，散布度，有用度に基づき語彙をランキングし，日本語教育的観点からの補正を加えて語彙表の語彙を選定し，レベル分けした。レベル分けと各レベルの語数は，データにおける語彙の出現頻度と分布の傾向を分析して設定し，各レベル 2000 語ずつの区切りで，2000 語レベル，4000 語レベル，6000 語レベル，8000 語レベル，1 万語レベルの 5 レベルを選定した。語彙は原則的に有用度指標順にランキングし，単語親密度でランク調整する方法で選定した。ただし，2000 語レベルの基本語彙に関しては日本語教育的観点からの補正を行い，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は単語親密度に関わらず 2000 語レベルに残した。

第 5 章では，語彙表の評価を行うため，語彙のテキストカバー率調査を行った。テキストカバー率調査には，日本語教育用に加工された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・エッセイを使用した。また，日本人向けに書かれた一般のテキストでは，日本語学習者が読む可能性のあるものとして，新聞，小説，Web サイトのテキスト，話し言葉コーパスを使用した。これらはそれぞれ「出題基準」の語彙のカバー率とも比較を行った。その結果，ほぼ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のほうが高いテキストカバー率を示した。また，本研究の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。

本研究の語彙表作成の手法や目的は先行研究において例のないものである。本研究の語彙表の利用可能性としては，教材利用，テスト利用，研究利用がある。今後の課題として，複合語の抽出，日本語教育的な表記の併記，日本語教育的観点からの補正のさらなる検討，中頻度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。