博士学位論文（東京外国語大学）

Doctoral Thesis (Tokyo University of Foreign Studies)

氏　　名	本田ゆかり
学位の種類	博士（学術）
学位記番号	博甲第 208 号
学位授与の日付	2016 年 3 月 12 日
学位授与大学	東京外国語大学
博士学位論文題目	大規模コーパスに基づく日本語教育語彙表の作成

Name	Honda, Yukari
Name of Degree	Doctor of Philosophy (Humanities)
Degree Number	Ko-no. 208
Date	March 12, 2016
Grantor	Tokyo University of Foreign Studies, JAPAN
Title of Doctoral Thesis	Creating a corpus-based educational word list for learners of Japanese

大規模コーパスに基づく日本語教育語彙表の作成

Creating a corpus-based educational word list for learners of Japanese

東京外国語大学大学院

地域文化研究科　博士後期課程

言語教育学講座

2015 年　9 月

本田ゆかり

目次

図表目次

第1章　はじめに

1.1節　本研究の目的

本研究は，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り定量化し，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，そして，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ことを主な目的とした基本語彙を選定することにある。日本語教育では，従来から語彙調査の結果を資料とし，専門家の主観的な判断や判定（以下，「専門家判定方式」）によって複数の語彙表（以下，「従来型」）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』（Balanced Corpus of Contemporary Written Japanese：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，その出現頻度には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直接影響する。そのため，利用するコーパスが作成しようとしている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が，そのような例は過去にない。また，コーパス準拠の語彙表であっても語のレベル分けは専門家判定方式で行われていたり，既存の従来型語彙表の成果に依存していたりすることもある。

従来型の語彙表の中でも，日本語教育の現場や研究目的で幅広く用いられてきたのが，『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）（以下，「出題基準」）の語彙セットである。しかし，「出題基準」は日本語能力試験のために作られたものであり，現状のような幅広い利用を想定して作成されたものではない。そのため，日本語能力試験問題作成以外の広範囲な利用には問題が生じることもある。李（2013）は，「出題基準」について，「1.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」という 3 点を問題点として指摘している。

その後，2009 年に国立国語研究所より BCCWJ モニター版が公開され，領域内公開版も一部のユーザーの間で利用可能になった。これらを利用して作られたのが，「日本語を読むための語彙データベース」（松下，2011）と「日本語教育語彙表」（李・砂川，2012）である。どちらも語彙選定にはコーパスの出現頻度や統計指標を利用している。

松下（2011）では，モニター版が使用されている。モニター版は「書籍」や「Yahoo!知恵袋」という限定的なジャンルのテキストで構成されている。このコーパス出現頻度に基づくということは，「書籍」や「Yahoo!知恵袋」の高頻度語彙および分布の安定した語彙を日本語教育においても「重要」とみなしているということである。しかし，日本語学習者は，新聞，雑誌など，これら以外の様々なジャンルのテキストを読むであろうし，もっと幅広いテキストジャンルにおいて高頻度に出現する語彙を重要語彙として優先的に学習していくべきであろう。しかし，松下（2011）では，このモニター版が日本語教育語彙表作成に適したコーパスであるかどうかを検証した記述は見当たらず，コーパス出現頻度や統計指標を使って客観的にレベル分け等を行っているものの，それが本当に妥当なものであるかどうかは疑問が残る。

一方，李・砂川（2012）はモニター版よりも幅広いジャンルのテキストを収録する領域内公開版を利用している。さらに，日本語教科書のデータも使っている。このように，李・砂川（2012）では，使用するコーパスが日本語教育語彙表作成という目的に合うものとなるような配慮がなされている。しかし，語彙選定やレベル分けの方法については，「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。」（李，2013，p.13）という考え方から専門家判定方式に依存し，その方法論は従来型に近い。

本研究では，語彙選定における主観的要素を可能な限り排除し，コーパスバランスについて詳細に検討したうえで，日本語の書き言葉を読んで理解するために必要な語彙を中心とした語彙リストを作成する。近年，インターネットの普及により世界中の

どの地域にいても日本語に触れることが可能になった。現在，日本語学習者は，国内で約 15.7 万人（2013 年，文化庁調べ），海外で約 399 万人（2012 年，国際交流基金調べ）である。国内学習者数よりも海外学習者数のほうが圧倒的に多いが，インターネットなどの媒体を利用すれば，学習リソースが限定される海外の学習者でも日本語を読む機会は得ることができる。そして，日本語のテキストを読む力を向上させるためには，多様な書き言葉テキストにおいて高頻度で出現し，分布の安定した語彙から学習するのが効果的である。しかし，既存の日本語教育語彙表にはそのような観点で作られた例がなく，日本語を読んで理解するための重要語彙は何かというニーズに応えるものがない。

日本語教育語彙表は,権威的,代表的なものが一つあればよいというものではない。それぞれの目的に応じて作られ，活用されるべきである。日本語の書き言葉を読んで理解するためには，書き言葉における重要語彙を選定した語彙表が役立つ。インターネットなどの利用しやすい媒体を通して日本語のテキストを読み，生の情報が得られれば，日本に馴染みのない国や地域における学習者も日本という異文化に対する理解を深めることができる。そして，それは世界各国に親日家を増やすという日本語教育の大きな目的にもつながる。

本研究で作成する日本語教育語彙表は，書き言葉を収集した大規模コーパスを用いて，日本語を読むことを中心としたときに重要な基本語彙を選定するものである。コーパスの利用に際しては，コーパスのバランスが本研究の日本語教育語彙表作成に適しているかどうかを十分に検討する。そして，語彙の選定は，語彙の重要度を統計指標によって定量化してランキングし，単語親密度（天野・近藤，1999）によって日本語教育的観点からの語彙の難易度を加味してそのランクを再配列する方法をとる。ただし，単語親密度は日本語教育的な難易度をそのまま反映するというものではない。そのため，単語親密度特有の影響で日本語教育的に基本的な語彙が下位のレベルに配置された場合，元のランクに戻すなどの調整も行う。このような方法で日本語の特に書き言葉の実情に合った基本語彙を選定しレベル分けを行う。本研究は既存の日本語

教育語彙表において不十分であった点の改善を試み，日本語教育語彙表開発研究のさらなる発展につなげようとするものである。

1.2節　構成

本研究の構成は以下の通りである。

第 1 章では，本研究の目的と構成について述べる。

第 2 章では，先行研究を概観する。まず，従来から作られてきた語彙調査に基づく日本語教育基本語彙の歴史についてまとめる。複数の語彙表が作られているので，それらの語彙表間の一致率に関する研究なども踏まえ，従来型の日本語教育語彙表の特徴と問題点についてまとめる。

次に，コーパスに基づく教育語彙表について見ていく。ここでは，コーパス準拠の語彙表開発が進んでいる英語教育の語彙表について先に触れる。そして，現在利用可能な日本語コーパスについて紹介し，日本語教育におけるコーパス準拠の語彙表の現状についてまとめる。

さらに，日本語コーパス研究の中で進められてきた語の単位と漢字表記の問題について見ていく。日本語は英語のように分かち書きの習慣がなく，どこまでを一語と認定するかは難しい問題である。また，表記の問題も複雑である。

第 2 章の最後には，先行研究についてまとめ，今後どのような日本語教育語彙表が必要とされるかという問題について考察する。

第 3 章では，研究方法について述べる。ここでは，まず，本研究で作成する語彙表のデザインや対象者，利用範囲などについて説明する。そして，語彙表作成に用いるコーパスの選定方法，および，サブコーパスバランスの調整方法について述べる。次に，コーパスを頻度リスト化し，散布度，有用度などの分析用基礎統計の算出方法と，単語親密度（天野・近藤，1999）の付与について述べる。さらに，語彙の選定とレベル分けの方法について説明する。

第 4 章では，コーパスに基づく日本語教育語彙表作成の実際について述べる。第 3 章で説明した方法で，コーパスの選定と再構築を実施し，分析用基礎統計を算出したうえで，語彙のランキングとレベル分けを行う。

第 5 章では，完成した語彙表を評価するためにテキストカバー率調査を行う。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，日本人向けに書かれた一般のテキストを利用する。

第 6 章では，本研究の結論と展望について述べる。本研究で作成する語彙表とその評価について結果をまとめ，考察する。そして，本研究の意義と，今後の課題として残された問題について記す。

第2章　先行研究の概観

第 2 章では，先行研究を概観する。2.1.では，従来から行われてきた，語彙調査に基づく日本語教育語彙表の歴史を振り返る。2.2.2.2 節 では，コーパスに基づく教育語彙表について扱う。まず，コーパス準拠の語彙表研究が進んでいる英語教育の語彙表についてまとめる。そして，2010 年ごろから進んできた日本語教育におけるコーパスに基づく語彙表の例を見ていく。2.3.では，日本語コーパス研究について，語の単位と表記に焦点を当て，まとめる。2.4.では，先行研究を踏まえて問題提起を行い，本研究の意義について述べる。

2.1節　従来の語彙調査に基づく日本語教育基本語彙

ここでは日本語教育語彙表の歴史を概観し，その特徴や問題点について述べる。2.1.1.では，日本語コーパス開発以前に盛んに行われた日本語の語彙調査についてまとめる。2.1.2.では，基本語彙の概念について定義する。2.1.3.では，過去に開発された主な日本語教育語彙表について見ていく。2.1.4.では，このような日本語教育語彙表間の一致率を調査し，基本語彙選定の妥当性を検証した研究について概観する。最後に 2.1.5.では，従来の日本語教育基本語彙および日本語教育語彙表の特徴と問題点について整理する。

2.1.1. 従来の日本語語彙調査

従来の日本語教育語彙表や日本語教育基本語彙と呼ばれるものは，語彙調査に基づいて作られたものが多い。現代日本語の語彙調査は 1950 年代ごろから国立国語研究所によって行われてきた。その種類は，新聞，雑誌，教科書，テレビ放送など多岐にわたる。これらの語彙調査の大きな目的の一つは，国語教育や日本語教育に役立つ基

本語彙の選定に基礎的な資料を提供することであり，これらをもとに様々な語彙表が作られた（表 1)。

表 1　国立国語研究所の現代語語彙調査

調査対象	資料の期間	調査方法	述べ語数	異なり語数	調査単位
① 新聞 1 か月	昭和 24 年 6 月	全数調査	24 万	1.5 万	β'
② 婦人雑誌	昭和 25 年	標本調査	15 万	2.7 万	α
③ 総合雑誌	昭和 28 年～29 年	標本調査	23 万	2.3 万	β
④ 雑誌九十種	昭和 31 年	標本調査	53 万	4.0 万	β
⑤ 新聞 3 紙	昭和 41 年	標本調査	300 万 200 万	21.3 万 －	短 長
⑥ 高校教科書	昭和 49 年	全数調査	59 万 45 万	1.6 万 4.1 万	W M
⑦ 中学校教科書	昭和 55 年	全数調査	25 万 20 万	0.8 万 1.8 万	W M
⑧ テレビ放送	平成元年 4～6 月	標本調査	14 万	2.6 万	長'
⑨ 雑誌 70 誌	平成 6 年	標本調査	105 万	4.8 万	B

（山崎（2009）の表をもとに，編集した）

表 1 の①は朝日新聞 1946 年 6 月の 1 か月分の調査である。初期のもので，標本調査ではなく全数調査が行われている。調査単位は後に開発される β 単位[1]に近いものである。②③④は語彙調査の発展期に当たり，この頃に調査方法や語彙調査のための理論である計量語彙論が確立した（山崎，2009)。③には，単位語，見出し語，使用度数，使用率，述べ語数，異なり語数などの基本的な用語が定義されている。また，日本語は分かち書きの習慣がないため，語の認定にあたっては揺れが生じやすいが，②③④を通じて β 単位が開発され，この問題が解消された（山崎，2009)。β 単位は⑤の短単位とほぼ同じものである。また，④には基本語彙の客観的な選定方法も収録されている。

その後，1965 年から語彙調査にコンピュータが導入され，調査語数が拡大した。調

[1] 語の単位（国立国語研究，1984, p.81）
β 単位（短単位とも言われる）　型紙/どおり/に/裁断/し/て/外出/着/を/作り/まし/た/。
長単位　型紙どおり/に/裁断し/て/外出着/を/作りました/。
M 単位（形態素 morpheme の略）　型/紙/どおり/に/裁断/し/て/外出/着/を/作り/まし/た/。
W 単位（語　word の略）　型紙どおり/に/裁断して/外出着/を/作りました/。

査単位も，短単位と長単位という 2 種類が併用されるようになった。⑥⑦は高校の教科書と中学校の教科書を対象とした全数調査である。⑧は話し言葉を取り上げた最初の調査で，テレビ局 6 局 7 チャンネルの音声と文字の両方を対象としている。このようにして語彙調査は進められ，調査方法や語の単位の認定方法なども進歩してきた。

2.1.2. 基本語彙とは

ここでは基本語彙の概念について定義する。先行研究では基本語彙を以下のように定義している。:

(1) 使用率が大きく，しかも対象とする言語作品あるいは言語体系の中にいくつかの層を設けて考えることができる場合（例えば，雑誌であれば，実用記事・文芸作品・趣味など掲載する内容別の層を設け，また平安時代物語であれば作品別に層を設けることができる），できるだけ多くの層にわたって出現する語の集合。(『国語学大辞典』，樺島，1980, p.143）

(2) ある目的のために語彙調査によって選定された使用率[2]が高く，使用範囲の広い語を選んだものを一般的に基本語彙という。（秋元，2002, p.37）

(3) 基本語彙は，ある目的のために頻度や重要性を考慮して選定されたものである。それぞれの分野で調査した言語資料を基に，頻度が高くて使用範囲が広く，当該分野で重要な語を客観的に選定している。（近藤・小森，2012, p.217）

つまり，基本語彙とは語彙調査に基づき，様々な分野に幅広く高頻度で出現する語彙である。これに対し，専門語彙は特定の分野に顕著に多く現れる語彙である。外国

---

[2] 頻度（使用度数）がデータの延べ語数の影響を受けるものであるため，それを相対化するために考案された指標。　語 w の使用率 $= \frac{\text{語 w の使用度数}}{\text{その文章の中の全ての語の使用度数}} \times 1000$（国立国語研究所（1984）p.83）

語教育では一般に 2000 語から 3000 語程度の基本語彙を学習した後，学習者の興味や学習目的に応じてある特定分野の専門語彙を指導すると学習効率が上がると言われている（中條，2009, p.10）。

一方，基本語彙と類似した概念に基礎語彙がある。基本語彙が頻度や使用範囲に基づき客観的に選定されるのに対し，基礎語彙は C.K. Ogden の Basic English の考え方に基づいたもので，人為的に選ばれた語彙である。Ogden は，人間の言語表現の内容を心理学的に分析した結果，850 語で日常生活のことがらは全て表現できるとし，これを基礎語彙とした。先行研究では以下のように定義されている。:

(1) それによって生活の大部分の必要をまかなうことができる語彙で，語を組み合わせることによって必要な意味を表すことができるように選ばれた語，あいさつなど対人接触に必要な語，質問や聞き返しなど情報を獲得するために必要な語，および人に質問して知ることが心理的に困難な，生理上の必要をみたすための語から成ることが考えられる。（『国語学大辞典』，樺島，1980, p.143）
(2) 日常の言語生活に必要な最小限の語を，一定の数だけ主観的な判断によって体系的に選定したもの。（秋元，2002, p.37）
(3) 特定の言語社会において生活し，知識，情報を得るための基盤となる語彙のこと。日常生活で使用される語彙の中でも共通性，頻度が極めて高いもの。（近藤・小森，2012, p.217）

すなわち，基礎語彙とは，日常の生活言語生活に必要な最小限の語を心理学的な見地から選定したものである。頻度や分布から客観的に選ばれる基本語彙に対して，基礎語彙は専門家の主観的判断によって選ばれるものであるので，類似概念ではあるが，はっきりとした違いがある。

このように言語教育で使われる語彙表は，基本語彙，専門語彙，基礎語彙などに分けることができる。ただし，日本語教育においては「基本語彙」と「基礎語彙」とい

う用語が必ずしも明確に使い分けられておらず(近藤・小森, 2012, p.217),「基本○○」という名称の語彙表であっても,基礎語彙の観点から選ばれた語彙であることもある。また，基本語彙とされていても判定の際に専門家の主観において，基礎語彙の影響から生活語彙や具体物を表す語彙などが「基本的」な語彙とされている可能性もないとはいえない。

基本語彙と基礎語彙は異なるものであるにもかかわらず，日本語教育において明確に使い分けられていないことは利用者側にも誤解を生み問題である。利用者は少なくともその語彙表がどのような語彙を集めたものなのかということを理解して教育や研究に使うべきである。

本研究で選定する語彙は基本語彙である。語彙選定は日常生活に必要かどうかなどの基礎語彙的な観点で行わず，コーパスの出現頻度と語彙分布の安定性を重視する。日本語教育ではこれまで多くの基礎語彙や専門家判定方式による基本語彙が選定されてきたが，専門家判定方式に頼らずに客観的に選定した基本語彙はほとんどない。一方，学習者を取り巻く環境は変化している。生活語彙に特別焦点を絞ることなく日本語の実際の使用を反映した語彙を選定することによって，生活場面に限らない様々な場面での日本語の理解につながり，学習者の多様なニーズにも応えることができると考える。

2.1.3. 日本語教育基本語彙の選定

ここでは主な日本語教育語彙表の歴史を振り返る。なお，本研究は外国語として日本語を学習する場合の日本語教育語彙表作成を目的としているので，ここでは，外国人日本語学習者を対象としたものと，日本人を対象とした国語教育の教育基本語彙とは区別する（表 2)。

表 2　語彙調査に基づいて作られた主な日本語教育基本語彙表

語彙表の名称（通称を含む）	出典	発行年	編著者名	収録語数	特記事項
「基礎日本語」	『基礎日本語』	1933	土居光知	1100	日本語を 1000 語に絞ってみるという発想から作られた語彙表。
「基本簡易ニッポン語」	『基本簡易ニッポン語』	1942	情報局	300	日本語を普及させるために作成された語彙表か？
日本語基本語彙	『日本語基本語彙』	1944	岡本禹一	2012	成人日本語学習者を対象に，「学習上の便宜として第一次の基礎になる語彙の標準とする」ものを，専門家の判定方式によって選定した。
日本語教育における基礎学習語	『日本語教育』2,4.5	1963～64	加藤彰彦	1393	語彙調査，日本語教科書，辞書の共通する語彙を調べた。
Practical Japanese-English Dictionary	Practical Japanese-English Dictionary	1970 1978	玉村文郎	3209	語彙調査に基づき，専門家の修正を加えて選定した。
留学生のための基本語彙表	大阪外国語大学『日本語・日本文化』2	1971	樺島忠夫，吉田弥寿夫	1803	高校教科書の語彙調査に基づき，留学生を対象に作成された語彙表。
外国人のための基本語用例辞典	『外国人のための基本語用例辞典』	1971, 1975	文化庁国語課	2500 3691	「日本語を学ぼうとするあらゆる外国人にとって，どうしても身に付けなければならないと思われる基本的な語を集め，その活用・意味・使い方などを開設する」もの。方法は，語彙調査を資料とした専門家判定方式。
A Classified List of Basic Japanese Vocabulary	A Classified List of Basic Japanese Vocabulary	1977	J.V. Neustupný	1796	モナシュ大学日本語学部入学試験受験者を対象とする。既存の教科書から入門期の基本語彙 1750 語を選定し，トピック別（18 分野，144 項目）に分類している。
日本語教育語彙資料 (1)(2)―低学年初級 500 語	『日本語教育語彙資料 (1)(2)―低学年初級 500 語』	1979	国立国語研究所	500	初級の学習者が最初に学ぶ生活語彙を偏りなく選定したもの。
日本語教育基本語彙七種 比較対照表	『日本語教育基本語彙七種 比較対照表』	1982	国立国語研究所	6073	「外国人に対する日本語教育を直接の目的として選定あるいは編集された，既存の語彙表七種について，それらに収録させた語彙を集め，それぞれの語彙項目について各語彙表間の異同を比較対照させたもの」
国語研教育基本語彙	『日本語教育のための基本語彙調査』	1984	国立国語研究所	6060	留学生等外国人の日本語学習者が，はじめに学習すべき日本語の一般的・基本的な語彙について妥当な標準を得ることを目的とする。『分類語彙表』を材料として用い，専門家による判定方式で，基本語の 6060 語のうち，より

					基本的なものを「基本語二千」(2030 語)としている。
日本語教育基本 2570 語	『NAFL Institute 日本語教師養成通信講座 8 日本語の語彙・意味』	1987	玉村文郎	2570	「日本で生活し,学習・研究・研修・業務などの活動をする外国人のための必修語彙として選定したもので,いわば<最小必修語彙>と目されるもの」
「初級日本語教科書によく使われる語」	『日本語教育機関におけるコース・デザイン』	1991	日本語教育学会	3121	大学進学の予備教育機関で使用している日本語の教科書 5 種に使用されている語彙を調査している。
品詞別・A~D レベル別 1 万語語彙分類集	『品詞別・A~D レベル別 1 万語語彙分類集』	1991 ※ 1998 年改訂	専門教育出版	10000	専門教育出版主催「日本語学力テスト」の基準を定めるための出題基準語彙。各種語彙調査から選定委員が選んだ語彙をレベル分けしている。
「簡約日本語」	『簡約日本語の創成と教材開発に関する研究』	1992	野元菊雄	1000 1000	「国際共通語としての日本語を世界により広く進める」ために「エッセンスとしての日本語を創り出す」ことを目的とする。
「出題基準」	『日本語能力試験出題基準』	1994 ※ 2002 年改訂	国際交流基金・日本語国際教育協会	10000	「日本語能力試験」の出題基準語彙。各種語彙調査から選定委員が選んだ 10000 語の語彙を 4 級から 1 級のレベルに区分している。
児童生徒に対する日本語教育のための基本語彙調査	『児童生徒に対する日本語教育のための基本語彙調査』	1999	工藤真由美	6001	外国人児童生徒が,日本の小中学校での教育を受けるにあたり,はじめに学習すべき日本語の妥当な標準を得ることを目的とする。国語研教育基本語彙の「基本語 2000」のほか,外国人児童向けの日本語教科書,子供向けの辞典などを含む 6 種の資料に基づき語彙調査を行い,語彙を選定している。
中級用語彙 —基本 4000 語	『日本語教育』116	2003	玉村文郎	4000	「日本語教育基本 2570 語」に「中級段階の必修語彙を想定して増補」したもの。

語彙頻度調査を実施してそこから語彙を選定する方法による国語教育の教育基本語彙表は,古くは戦前から作られている。坂本(1943)は『日本語基本語彙』の中で幼年児童読物について語彙調査を行い,「幼年基本語彙」を発表した。その後も,田中(1956)『学習基本語彙』や,池原(1957)『国語教育のための基礎語体系』,坂本(1958)『教育基本語彙』,坂本(1984)『新教育基本語彙』などが開発された。

外国人に対する日本語教育を目的とした基本語彙の選定も第二次世界大戦前から行

われている。以下，主な日本語教育基本語彙表について作成年順に記す。表 2 に見られる「基本」という名称の付いた日本語教育語彙表の中には，頻度や分布などに基づいて選定された厳密な意味での基本語彙ではなく基礎語彙の考え方で作られているものもあるが,ここでは日本語教育を目的とした語彙表作成の歴史を広く捉えたいので,基礎語彙の主なものも表中に含めた。

初期に作られたものには,「基礎日本語」(土居，1933),「基本簡易ニッポン語」(情報局，1942）などがある。これらは，戦中，諸外国において日本語を普及させる目的のもとで，必要最低限の語彙を抽出するという立場から作られたものである。このほか,成人日本語学習者を対象に,「学習上の便宜として第一次の基礎になる語彙の標準とする」ことを目的とした『日本語基本語彙』(岡本，1944）が発表された。これは辞典などを参考とし，専門家が語を認定する方法で選定を行っている。

1960 年代に入ると，加藤（1963～64)「日本語教育における基礎学習語」が発表された。これは，坂本（1958)『教育基本語彙』や岡本（1944)『日本語教育基本語彙』などの語彙調査や日本語教科書および辞典などの 6 種の資料について共通して出現する語彙を調査したものである。1970 年代には，国内外で日本語教育が広く行われるようになり，日本語学習者の数も増加する傾向にあった。

その後，語彙調査ではコンピュータが導入され，この成果を利用した日本語教育語彙表の開発も進んだ。樺島・吉田（1971)「留学生のための基本語彙表」は日本に滞在する外国人留学生を対象とし，高校教科書の語彙調査に基づいて 1803 語の基本語彙を選定している。また，学習者のレベルを考慮した語彙表も作られるようになる。国立国語研究所（1979)「日本語教育語彙資料(1)(2)—低学年初級 500」は，初級学習者が最初に学ぶ生活語彙を，わずか 500 語であるが偏りなく選定したものである。文化庁国語課 (1971)『外国人のための基本語用例辞典』や, 玉村 (1970, 1978) *Practical Japanese－English Dictionary* のように，辞典として発行されたものもある。これらも語彙調査を資料とし，専門家が語を選定するという方法を採っている。そのほか, J.V. Neustupný (1977) があるが，これはモナシュ大学日本語学部入学試験受験者を

対象としたもので,既存の教科書から入門期の語彙1750語を選定した語彙表である。この時代あたりから，対象とする学習者やそのレベルや目的などに考慮した語彙表も増えてくる。

1980年代に入ると，国立国語研究所による特別研究「日本語教育のための基本的な語彙に関する調査研究」並びに「比較対照研究」の成果として,『日本語教育基本語彙七種 比較対照表』と,『日本語教育のための基本語彙調査』が発表された。『日本語教育基本語彙七種 比較対照表』は,「外国人に対する日本語教育を直接の目的として選定あるいは編集された既存の語彙表 7 種について，それらに収録された語彙を集め，それぞれの語彙項目について比較対照させたもの」である。『日本語教育のための基本語彙調査』は「留学生等外国人の日本語学習者が，専門領域の研究または職業訓練に入る基礎としてはじめに学習すべき日本語の一般的・基本的な語彙についての妥当な標準を得る」という目的で行われた。ここでは「基本語六千（6060 語)」を選定し，そのうち，より基本的な 2030 語を「基本語二千」としている。選定方法は，日本語のシソーラスである『分類語彙表』(国語研究所，1964）を基本度判定の材料として用いる，専門家判定方式が採用されている。これは,『分類語彙表』所載のそれぞれの語に対する専門家の「基本度意識」の調査である。『分類語彙表』には，国立国語研究所(1962)『現代雑誌九十種の用字用語』語彙表所載の高使用率語と，坂本一郎(1958)『教育基本語彙』[3]所載の 22500 語が収録されている。このような語彙を専門家判定方式で選定している「国語研基本語彙表」は，語彙調査をベースとして選定された語彙を含む『分類語彙表』収録語から選ばれているものの，やはり専門家の主観的判断によって選ばれた語彙と言える。

このほか,1980年代に作られた主な語彙表には玉村(1987)「日本語教育基本 2750」がある。これは,「日本で生活し，研究・研修・業務などの活動をする外国人のための必修語彙として選定したもので，いわば＜最小必修語彙＞と目されるもの」であるとされている。詳しい選定方法などの説明がないものの，専門家が周到に企画し，慎重

[3] 小・中学校用の教育語彙表で，国語教育専門家の判断を得点化して選定している。

に選定した語彙であるとされ（飛田・佐藤，2002），基礎語彙の考え方に基づいたものだと考えられている（秋元，2002)。また，2003 年には，中級段階の必修語彙を想定して増補した玉村（2003)「中級用語彙―基本 4000」が発表されている。

1990 年代に入ると，日本語教育語彙表には多様化の傾向が見られる。語彙表に選ばれる総語数が 1 万語にまで増え，学習者の習熟度に合わせてレベル分けされた語彙表も複数作られた。中でも日本語教育の現場や研究で支持され，幅広く利用されているのが，国際交流基金・日本国際教育支援協会（1994)『日本語能力試験出題基準』（※2007 年改訂）に収録されている語彙表である。「出題基準」は，国際交流基金および日本国際教育支援協会が主催する日本語能力試験[4]の問題作成者向けに作成されたもので，試験問題の基準となる語彙や文法項目の全てではないが大部分が公表されている。

表 3　『日本語能力試験出題基準』語彙のレベル別語数

レベル（級）	各レベルの語数	累積語数
4 級	800	800
3 級	700	1,500
2 級	4,500	6,000
1 級	4,000	10,000

表 4　日本語能力試験（2010 年改訂前）の認定基準

レベル	認定基準
1 級	高度の文法・漢字（2,000 字程度)・語彙（10,000 語程度）を習得し，社会生活をする上で必要な，総合的な日本語能力（日本語を 900 時間程度学習したレベル）
2 級	やや高度の文法・漢字（1,000 字程度)・語彙（6,000 語程度）を習得し，一般的なことがらについて，会話ができ，読み書きできる能力（日本語を 600 時間程度学習し，中級日本語コースを修了したレベル）
3 級	基本的な文法・漢字（300 字程度)・語彙（1500 語程度）を習得し，日常生活に役立つ会話ができ，簡単な文章が読み書きできる能力（日本語を 300 時間程度学習し，初級日本語コースを修了したレベル）
4 級	初歩的な文法・漢字（100 字程度)・語彙（800 語程度）を習得し，簡単な会話ができ，平易な文，又は短い文章が読み書きできる能力（日本語を 150 時間程度学習し，初級日本語コース前半を修了したレベル）

（「日本語能力試験 JLPT」http://www.jlpt.jp/）

[4] 日本語能力試験は 2010 年に大幅な改訂が行われたが，改訂後の出題語彙は公表されていない。

「出題基準」のレベルは，難度の高い順に 1 級，2 級，3 級，4 級の 4 つの級に分かれている。各級の語数は等間隔ではない。4 級が 800 語，3 級が 700 語であるのに対し，2 級が 4500 語，1 級が 4000 語と急に語数が増え，特に，3 級と 2 級の間に大きな差があるのが特徴である。レベルについての説明は，3 級が「日本語を 300 時間程度学習し，初級コースを修了した程度」，2 級が「日本語を 600 時間程度学習し，中級コースを修了した程度」とあるが（表 4），日本語教育において「出題基準」が最も信頼性の高い語彙表であるとすれば，語彙習得上の数の目安として，初級で 1500 語程度，中級で 6000 語程度と考えられているとみてよいであろう。

3，4 級と 1，2 級では，語彙の選定方法も異なっている。3，4 級の語彙は，当時の調査により使用機関数が多いとされた 11 種類の初級用日本語教科書を基礎資料とし，日本語教育に関する語彙調査（国立国語研究所編『日本語教育基本語彙第 1 次資料―上位 2000 語』），および J.V. Neustupný (1977) *"A Classified List of Basic Japanese Vocabulary"* を参考資料として選定されている。つまり，3，4 級の語彙は，語彙調査等を参考とし，日本語教科書という限定したジャンルにおける使用域に基づいて選定されたものである。

一方，1，2 級の語彙では，中上級用の日本語教科書の語彙から選定するという方法は取られず，語彙調査の資料をもとに日本語教育の立場から修正を加えるという方法で選ばれている。中・上級用教科書は，既成の小説・随筆・論文・記事などを引用，編集したもので構成されていることが多い。すなわち，中・上級用教科書の語彙は，語彙教育の立場から検討して選ばれたものではないので，トピックに依存したものが多い。このような理由によって，1，2 級では日本語教科書が基礎資料として用いられていない（表 5）。しかし，いずれも語彙調査資料によって候補語を出し，専門家判定方式で語彙を認定していくという方法論においては同じである。

表 5　1，2 級「出題基準」語彙の選定資料

語彙調査	採録語彙数
国立国語研究所編（1984)『日本語教育のための基本語彙調査』	6,060
国立国語研究所編（1980)『日本人の知識階層における話し言葉の実態』	5,341
「3，4 級出題基準」作成のための提出語彙調査によって得られた語	4,487
外国人の日本語能力に関する調査研究協力者会議（1982）『外国人留学生の日本語能力の標準と測定に関する調査研究について』	5,167
国立国語研究所編（1964)『分類語彙表』	32,600
国立国語研究所編（1987)『中学校教科書の語彙調査Ⅱ』(※使用度数上位 3,290 語のみ)	3,290
国立国語研究所編（1984)『高校教科書の語彙調査Ⅱ』(※使用度数上位 3,067 語のみ)	3,067
合計	60,012

このほか，1990 年代に作られた類似の語彙表には，専門教育出版（1991)『品詞別 A~D レベル 1 万語』(※1998 年改訂）に収録された語彙表がある。これは，専門教育出版主催の「日本語学力テスト」の語彙レベルを定めるために作成された語彙表で，「出題基準」と同様，語彙調査資料をベースとした専門家判定方式で語彙の選定が行われている。語彙レベルは A~D の 4 段階に分けられている。ここで用いられている語彙調査資料は，国立国語研究所（1984)『日本語教育のための基本語彙調査』，国立国語研究所（1964)『分類語彙表』や，国語辞典などを含む 6 種である。

90 年代には，このような試験問題作成の基準となる語彙表の作成が進んだほか，これ以前にも作られてきた限られた基本語彙で表現することを目的とするタイプの，いわゆる「表現の語彙表」も継続して作られている。野元（1992）の『簡約日本語の創成と教材開発に関する研究』は，土居（1933)『基礎日本語』や情報局（1942)『簡約基本ニッポン語』などの流れをくむもので,「国際共通語としての日本語を世界により広く進める」ために「エッセンスとしての日本語を創り出す」という立場で作られている。第一次で 1000 語，第二次で 1000 語の合計 2000 語が選定されている。

そのほか，日本語学習者が多様化したことを背景に，特定の目的の日本語教科書などを調査して作られた語彙表も作られた。日本語教育学会（1991)「日本語の初級教

科書によく使われる語」は，大学進学の予備教育機関で使用されている日本語教科書5種[5]に使用されている用語を調査してまとめたものである。また，工藤（1999）『児童生徒に対する日本語教育のための基本語彙調査』は,「外国人児童生徒が日本語の小中学校（特に，小学校）での教育を受けるにあたって，初めに学習すべき日本語の基本的な語彙についての妥当な標準を得る」ことを目的として作られた語彙表である。資料は『日本語教育のための基本語彙調査』「基本二千」や「簡約日本語語彙表」のほか，子供向けの辞典や日本語教科書などの6種を利用し，これらの見出し語の共通度を調査した結果を発表している。

このように，戦前から数多くの日本語語彙表が作られており，時代に伴い日本語教育の目的が変化し，対象者も多様化し，それに合う語彙選定の方法も試行錯誤が繰り返されてきた。初期には1000語程度の小規模な「基本語彙」が選ばれることが多かったが，1990年代には選定される語数は最大1万語にまで拡大している。これはコンピュータの発達により大規模データも容易に処理できるようになったという背景も影響していると考えられる。

これまでの日本語教育語彙表の作られ方には，大きく分けて三つのタイプがある。一つは，学習者にとって必要最低限の語彙を選ぶタイプのものである。これには,「基礎語彙」タイプのものも含まれるが，選定方法は語彙調査などの統計資料に頼らずに専門家の目で判定することに重きを置いている。そのため，主観的に選ばれた語彙であると言える。もう一つは，既存の語彙調査や日本語教科書の語や辞書の見出し語などの重なりを見て共通する語を抽出し，それを語彙表とするものである。これらは基本的に重なりだけを見ており，頻度による重みづけなどはなされていない。最後は，語彙調査資料を参考に語を選定しようとするものである。しかし,これらについても，最終的には専門家の判定によって語の認定やレベル分けを行っているタイプのものが多い。つまり，語彙調査資料を参考にしていても基本的にはこれらも専門家判定方式

[5] 国際学友会日本語学校（1977）『日本語Ⅰ，Ⅱ』，言語文化研究所附属東京日本語学校(1988~1990)『COMMUNICATION JAPANESE STYLE Ⅰ・Ⅱ』，山田あき子（1990）『楽しく学ぶ日本語』インターカルト日本語学校，東京外国語大学附属日本語学校（1983）『日本語Ⅰ』，文化外国語専門学校（1989）『文化初級日本語Ⅰ・Ⅱ』文化外国語専門学校

によるものである。このように長い歴史の中で，専門家判定方式による日本語教育語彙表の開発は，すでに十分に行われてきたと言える。

また，日本語教科書や辞書の見出し語などを用いて，その重なりを調査し重要語を選定しようとする試みなど，語彙選定の客観性を重要視する動きもあるが，その元となる語彙調査や日本語教科書の語彙選定は恣意的であったり，調査するデータのバランスが偏っていたりするという問題がある。

一方，コーパス頻度や統計指標に基づいて客観的に語彙を選定するタイプの日本語語彙表については，近年開発が始められた段階でありまだ数が少なく，今後はその成果が期待される。時代とともに学習者を取り巻く環境は変化し，語彙表に対するニーズも多様化している。特に，90 年代後半からはインターネットが社会基盤となり，大量の言語資料にアクセスできるようになった。日本語教育では古くから基礎語彙の観点から選ばれた生活語彙などを重要語とする語彙表などが専門家判定方式で作られ優れた成果を残したが，そのような枠組みを超え，実際に使用されている多様な日本語を理解するために有用な基本語彙を示すことも必要であると考える。

2.1.4. 語彙表間の一致率

語彙調査の結果を利用し，専門家判定方式によって選定された基本語彙はそれぞれどの程度一致しているのだろうか。同じような目的の下,「基本語彙」として選ばれたものであれば,各語彙表の語彙も大部分が一致するのではないかと期待したくなるが,それぞれの語彙表に収録されている語彙の一致度はそれほど高くない。

国立国語研究所（1984)『日本語教育のための基本語彙調査』(以下,「基本調査」)によれば,「基本調査」で選ばれた「基本語二千」「基本語六千」と，以下の 6 種の語彙リストとの間の一致度が非常に低い（表 6)。他の 6 種とは以下のもので，これらは語彙調査に基づき専門家判定方式で選定されたものである。:

(1) 岡本禹一（1944）『日本語基本語彙』（国際文化振興会）2012 語

(2) 加藤彰彦（1963, 64）『日本語教育における基礎学習語』（『日本語教育』第 2 号，および 4・5 号合併号，日本語教育学会）1393 語

(3) 玉村文郎（1970, 78）*Practical Japanese-English Dictionary*（海外技術者研修協会）3209 語

(4) 樺島忠夫・吉田弥寿夫（1971）「留学生のための基本語彙表」（『日本語・日本文化』第 2 号，大阪外国語大学留学生別科）1803 語

(5) 文化庁国語課（1971, 1975）『外国人のための基本語用例辞典』3691 語

(6) J.V. Neustupný (1977) *"A Classified List of Basic Japanese Vocabulary"*1796 語

「基本調査」によると，基本語二千を含む 7 種に共通する語は 285 語で全体の 14%にすぎない。また，他 6 種が 1393 語から 3691 語であるのに対し，比較する語彙表の語数が多すぎるため目安ではあるが，基本語六千を含む 7 種に共通する語は，全体の 4.7%にまで下がる。

表 6　「基本語二千」「基本語六千」の各語彙表間での共通度

	基本二千			基本六千		
	共通語数	累積	割合	共通語数	累積	割合
七種共通	285	285	14.0%	286	285	4.7%
六　〃	579	864	42.6%	585	864	14.4%
五　〃	435	1299	64.0%	513	1299	22.8%
四　〃	321	1620	79.8%	577	1620	32.4%
三　〃	248	1868	92.0%	799	1868	45.6%
二　〃	128	1996	98.3%	1331	1996	67.5%
一種のみ	34	2030	100.0%	1969	2030	100.0%

（国立国語研究所（1984, p.28）の表をもとに編集した）

これら 6 種の語彙表は，作成の目的，対象者，作成方法，作成年などが異なっている。(1)は 1944 年に作成されているが，(3)(4)(5)は(6)1970 年代に作成されたものである。さらに，(4)(5)は日本国内の留学生や外国人に向けたものであるのに対し，(6)は

オーストラリアの日本語学習者を対象としている。このように，(1)から(6)の語彙表は，同じ時代に作られたものばかりではなく，対象者も異なっている。特に名詞は，語彙表の語彙の最も大きな部分を占めているが，時代の変化を受けやすく，対象者によっても必要とされる語彙の種類に特色が出やすい。このようなことも全体の一致率が非常に低いことの原因であろう。

饗場（2011）は「基本調査」（6103 語），国立国語研究所（1982）『日本語教育基本語彙七種 比較対象表』（以下，「七種対照」）（6195 語）と，玉村文郎（2003）「中級用語彙―基本 4000―」（4043 語）（以下，「基本四千」）の 3 種に共通する語彙を分析している。まず，語彙表間の一致度を見たところ，3 種に共通する語彙は 3175 語であった。表 7 はその品詞別割合を示すものである。例えば，「基本調査」の名詞の「共通語彙の割合」（48.8%）は「基本調査」の名詞（4346 語）に占める共通語彙の名詞（2122 語）の割合を示している。

表 7　各語彙における共通語彙の品詞別割合

	共通語彙	基本調査		七種対照		基本四千			
品詞	語数	語数	共通語彙の割合	語数	共通語彙の割合	語数	共通語彙の割合	語数（6150 語で相対化）	共通語彙の割合
名詞	2122	4346	48.8%	4010	52.9%	2750	77.2%	4183	50.7%
代名詞	20	25	80.0%	37	54.1%	26	76.9%	40	50.6%
動詞	632	967	65.4%	1132	55.8%	715	88.4%	1088	58.1%
形容詞・形容動詞	171	312	54.8%	289	59.2%	189	90.5%	287	59.5%
副詞	125	255	49.0%	282	44.3%	143	87.4%	218	57.5%
連体詞	15	24	62.5%	25	60.0%	18	83.3%	27	54.8%
接続詞	24	35	68.6%	43	55.8%	25	96.0%	38	63.1%
その他	66	139	47.5%	377	17.5%	177	37.3%	269	24.5%
合計	3175	6103	52.0%	6195	51.3%	4043	78.5%	6150	51.6%

（饗場（2011）の表を編集した。「基本四千」の「語数」6150 語への相対化は筆者による。）

それぞれの語彙に占める共通語彙の割合は，「基本調査」が 52.0%，「七種対照」が

51.3%,「基本四千」が 78.5%で，一見,「基本四千」が他の 2 種と比べて高いように見える。しかし，これには「基本四千」の語数が他と比べて少ないことが関係している。仮に，基本四千の語数を他の 2 種と近い値（6150 語）に相対化して計算してみると，51.6%となり，他の 2 種と近い割合になる。

品詞別に見ると，名詞の共通語彙の割合がやや低い。それと比較すると，動詞をはじめ，形容詞，連体詞，接続詞がやや高くなっている。

さらに饗場（2011）は，対象の異なる語彙表との比較を行うため，上記 3 種の語彙表に,『児童生徒に対する日本語教育のための基本語彙調査』（6050 語）と『児童生徒に対する日本語教育のための語彙調査Ⅱ』を加えて，5 種の語彙表を比較した。その結果，名詞の共通部分の割合が，動詞と比べて大きく下回ることを明らかにした。このことから饗場（2011）は,「それぞれの語彙表を特徴づけている大きな要素の一つは，名詞の選定であることが推測される」と考察している。また，上記 3 種のうち 1 種の語彙表にしか入っていない語には複合語が多く見られることから，基本語彙間の共通度には複合語をどこまで入れるかが大きく関わっていると指摘している。

このように,各語彙表間の一致度は高くない。「七種対照」の比較に比べ,饗場（2011）の 3 種の語彙表の一致度は高く出ているが，これは対象とする学習者がほぼ一致していることや，比べている語彙表の種類が少ないことが関係していると考えられる。それでも，同じタイプの学習者（大学生以上の成人）を想定し，同じ目的で基本語彙を選んでいるにも関わらず，わずか 3 種類でも半分程度しか一致していないのは，かなり低いものと見てよいであろう。したがって，この結果だけ見ても，日本語学習者にとって何が基本語彙であるかという概念は非常に曖昧なものであると考えられる。

2.1.5. 日本語教育語彙表の特徴と問題点

ここまで語彙調査に基づく日本語教育語彙表について見てきた。ここからはこのような日本語教育語彙表の特徴と問題点をまとめる。従来の語彙調査に基づく日本語教

育語彙表の特徴には，以下の(1)～(4)が挙げられる。:

(1) 限定的なジャンルを対象とした語彙調査が語彙表作成の元データになっている。（2.1.3.参照）

(2) 日本語教科書だけを主なデータとして用いているものもある。（2.1.3.参照）

(3) 語彙調査を利用していても最終的な選定は専門家判定方式で行われている。(2.1.3.参照）

(4) 各語彙表間の一致率が低い。（2.1.4.参照）

まず，(1)については，語彙調査が行われたジャンルは，新聞，雑誌，学校教科書，テレビ放送である（2.1.1.参照)。しかし，頻度データをそのまま利用すると，テキストジャンルの語彙的特徴を受けることになる。過去の語彙調査に対して,「語彙調査の結果得られる語彙は，調査したデータの範囲を出ない。そのうえ，データの内容やテーマの偏りが，かなり頻度の高い語群にまで影響を及ぼしてしまうという点も見逃せない。」（北原，2005, p.20）という指摘もある。

(2)についても(1)と同様，基本語彙を抽出するのに，ある特定のジャンルの語彙的特徴を受けることの問題がある。初級レベルの学習者に向けた語彙を抽出するのに日本語教科書を利用するのは妥当であるが，教科書から抽出した語彙のみでよいかどうかは議論の余地がある。また,日本語教科書間の語彙の一致率もそれほど高くない（日本語教育学会，1991, p.73[6])。

(3)については，語彙調査に基づいた日本語教育語彙表であっても，最終的には専門家の判定方式を利用する語彙表が少なくなかった。また，日本語教育語彙表の歴史において，狭義の「基本語彙」よりも，そもそも語彙調査などの統計データに基づくことを前提としない「基礎語彙」や，限られた語彙で表現することを目的として，専門家の主観的判断で必要最低限の語彙を選ぶタイプの語彙表も数多く作られてきた。ま

[6] 5種類の初級日本語教科書の語彙の重なりを調査したところ，5種類に共通の語が 11.9%，4種類に共通の語が 10.8%，3種類に共通の語が 10.8%，2種類に共通の語が 18.8%，1種類のみのものが 47.7%という結果になった。

た，これは日本語教育に限られたことではない。英語教育でも 1980 年代から 1990 年代には，教師の直観・主観を優先する立場が，語彙表開発の基軸であった（石川，2008, p.162）。しかし，こうした主観的な語の調整は，語彙表の教育的妥当性を高める上で効果があったとしても，データ処理の客観性を保つことはできない。その結果，(4)のように語彙表間の一致率が低いという問題も生じてくると考えられる。

2.2節　コーパスに基づく教育語彙表

2.2.ではコーパスに基づく基本語彙の選定について概観する。2.2.1.では，コーパスという用語の定義を行う。2.2.2.では，コーパス言語研究が最も進んでいる英語教育分野においてどのようなコーパス準拠の教育語彙表が作成されてきたかをまとめる。2.2.3.では，現代日本語コーパスについて整理する。2.2.4.では，日本語コーパスを使った語彙表研究の現状をまとめ，今後の可能性について考察する。

2.2.1. コーパスとは

コーパス（corpus）とは，単に言語資料を集めただけのものではない。石川（2008）は著名なコーパス研究者による以下の(1)～(3)の定義に基づき，コーパスを定義している（以下，石川，2008，p.6 より引用）。:

(1) コーパスとは，言語の状態や多様性を特徴づけるべく選ばれた，自然に生起する言語テクストの集成である。(Sinclair, 1991, p.171)
(2) コーパスとは，何らかの言語，方言，または言語学の分析に用いられるその他の言語の下位区分を代表するとみなされるテクストの集成である。(Francis, 1982, p.7)
(3) コンピュータ・コーパスというのは，コンピュータに保存された大量のテクスト

のことであり，それ自体ではとりわけ刺激的なものではない。(Leech, 1992, p.106)

Sinclair はコーパスが自然言語の収集である点を強調し，Francis はコーパスが対象とする言語の全体を代表していることが条件とし，Leech はデータがコンピュータに記録されていることを基本要件としている。これらの見解を踏まえ，石川（2008, p.6）はコーパスを(4)のように定義している。:

(4) 言語研究におけるコーパスとは,自然な言語データをバランスよく電子的に集めたものである。

日本語学においては前川（2009, p.6）が次のようにコーパスを定義している。:

(5) 言語研究のための大規模なデータ。対象とする言語において実際に用いられた用例を,その言語の実情を正確に反映するように組織的に収集して,公開したもの。通常コンピュータで利用する。品詞情報などの検索用情報を付加したものが多い。

前川（2009）の定義も，コーパスとは自然言語の収集であり，その言語の全体を代表するものであり,電子化されたものであるという三つの要素が条件になっている点で,石川（2008）の定義と類似している。一方，Leech（1992）や前川（2009）は，データが大規模であることにも着目している。しかし，具体的に何語以上が「大規模」といえるのかを規定することは難しい。例えば初期にできた Brown Corpus は，100 万語でも作成された当初は「大規模」なデータであった。しかし，現在は様々な言語で数億語規模のコーパスがいくつも作られており,これらと同様に Brown Corpus が「大規模」と言えるかどうかは曖昧である。さらに，今後もコーパス規模の拡大傾向は進むと見られている。そこで,コーパスが大規模なデータであることは前提ではあるが,

本研究では先に述べた三点のみを基本要件とし，言語研究におけるコーパスとは，ある言語の実情を反映するように，バランスよく自然言語を収集し，電子化したものと規定する。

「言語の実情を反映するように，バランスよく自然言語を収集」したコーパスとはすなわち均衡コーパス（balanced corpus）のことである。前川（2013, p.14）はコーパスの均衡性について,「自然言語には多数の変種が存在する。話し言葉と書き言葉がその代表だが，ほかに媒体による差，使用場面や目的による差，性差，年齢差，地域差なども変種である。」としたうえで,「積極的に多数の変種をカバーして，対象言語の全体像を把握しようとするコーパスは均衡コーパスと呼ばれる。」と説明している。

コーパスの言語教育への応用で最も進んでいるのは英語教育である。英語教育では，コーパスを使った言語研究が 1964 年に Brown Corpus の公開以来進められ，基本語彙の選定や辞書の作成などにもコーパスが使われている。一方，日本語では，コーパスを使った言語研究自体が比較的新しいものである。そこで，まず 2.2.2.で英語コーパスに基づく基本語彙について概観した後，2.2.4.で日本語コーパスに基づく日本語教育基本語彙について見ていく。

なお，言語の種類に関わらず，言語教育に用いられる語彙表は基本語彙と専門語彙に大別できる。基本語彙とは「どの分野にも広く出現する語彙」のことで，専門語彙とは「ある特定の専門分野に顕著に表れる語彙」のことである（中條，2009）。教育現場では，その目的に応じて，基本語彙や専門語彙のリストがどちらも利用されている。本研究はコーパスに基づいて基本語彙を選定することを目的としているので，以下，基本語彙のリストについて考察していく。

2.2.2. コーパスに基づく英語教育語彙表

2.2.2.1. 英語教育におけるコーパス準拠の語彙表の語の単位

英語教育におけるコーパス準拠の語彙表について見ていく前に，まず語彙表の見出

し語（head word）に使われている語はどのような単位であるかを確認する。語彙表によって採用している語の単位が異なり，それは語彙表の語数にも関係している。

語彙表の単語は，主にレマ（lemma）を単位とするものと，ワードファミリー（word family）を単位とするものに分けられる。レマとは活用形や綴り字の違いを問わず，語幹と語類を同じくする各種の表記形を包含する基準形（canonical form）のことである（石川，2008，p.78）。したがって，一つのレマの下に含まれる単語は全て同じ品詞である（Nation，2001，p.7，Francis and Kučera，1982，p.486 より引用）。Thorndike and Lorge（1994）の頻度表で語数を数える基本単位としてレマが用いられたのをはじめ，コーパス言語研究ではレマ化（lemmatization）されたリストを使うのが一般的である。レマ化とは，綴りに着目した語の単位である表記形（word-form）をレマ単位にまとめることである。また，レマが辞書の見出し単位になりやすいことから，ほぼ同じ意味で見出し語という用語が使われることもある（石川，2008，p.78）。

単語を数える単位としてレマが使われるのは，それが学習負荷（learning burden）の重さを表しているためである（Swenson and West，1934）。学習負荷とは，ある項目（単語）を学習するために必要な努力量のことである。例えば，mend を知っている学習者が mends を学習する負荷は無視することができる（Nation，2001，pp.7-8）。レマは単語学習に関わる学習負荷を数えるのに適した単位である。

単語を数えるもう一つの単位はワードファミリーである。ワードファミリーとは，見出し語とその活用形および密接に関連する派生語から成る（Nation，2001，p.8）。体系的に使用される接辞は多く，基底語を知っている場合それらの接辞が作る派生語の学習負荷は軽くなる。このようなことをカバーできる点で，ワードファミリーは語彙表の語彙を数える単位として優れている。しかし，ワードファミリーに何を含み，何を含まないかを決定することは難しい。また，学習者の接辞の知識を前提にワードファミリーを規定しても，学習者の習熟度レベルがそれに達していないこともあり得るので，ワードファミリーの尺度も設定しなければならない（Nation，2001，p.8）。

語彙表の語を数える単位としてレマもワードファミリーも使用されているが，同じ

内容を指していても，レマはワードファミリーよりも語数が多くなるので注意が必要である。さらに，語彙表を作る際にはその目的に合わせていずれかの単位が使用されるべきである。また語彙表を利用する側も，その語彙表にはどのような単位が使われているのかを理解していなければならない。

2.2.2.2. コーパスに基づく英語教育語彙表

英語教育におけるコーパスに基づく基本語彙の主なものには，General Service List (West, 1953, 以下, GSL), 大学英語教育学会基本語リスト（大学英語教育学会, 2003, 以下, JACET 8000)，レベル別語彙リスト SVL 12000（アルク，以下，SVL 12000），などがある。これらはコーパスから出現頻度の高い語彙を選定するという原則に基づいて作られている。

GSL（West, 1953）は，英語教育のために選定された 2000 語の基本語彙で，語の単位はワードファミリーである。ワードファミリーは West の定めた基準によってまとめられている。2000 語は Head Words で，ワードファミリーを代表する語である。この 2000 語は 500 万語の書き言葉データからの頻度と range に基づいて選定されている。また，語の意味別にも頻度情報が付けられている。GSL は新しい語彙表ではないが，今も優れた語彙表として，辞書や教材の作成にも幅広く利用されている。

JACET 8000（大学英語教育学会基本語改定委員会，2003）は大学英語教育学会基本語改定委員会が BNC をもとに「日本の英語教育の現状を配慮して」選定した 8000 語である。レベルは 1000 語ごとに 8 レベルに分かれている。この語彙表はコーパスに基づき，統計的処理によって語彙を選出するという徹底した客観的，科学的方法において作成されている点において画期的で優れた語彙表である。本研究もコーパスに基づく客観的選定による語彙表の作成を目的としているため，JACET 8000 の作成方法から得られる知見は大きい。そこで，JACET 8000 の作成方法についてもう少し詳しく見ていく。

JACET 8000 は，まず，BNC から基準データを作成し，それとサブコーパスデータと照合して 8000 語選び出して順位を決め，さらに 8000 語の順位を教育的観点から再調整するという方法で作られている。基準データは，BNC から選ばれた「BNC 頻度順 100000 語リスト」と，変化形を基本形に寄せて上位語をリストした「BNC5516 語リスト」の二つである。サブコーパスデータとは，検定教科書，雑誌，新聞，映画，児童文学，BBC-CNN などのスクリプト，センター試験・STEP・TOEFL・TOEIC などの各種試験などから作成した，語彙出現頻度データ（8000 語）である。この BNC より作られた基準データとサブコーパスデータでの頻度の比較，および，対数尤度（log-likelihood）という視点からの比較を行い，8000 語の順位調整をしている。さらに，上位 3000 語については，高校教科書コーパス順位との照合により，順位が再調整されている。これは，時事用語や俗語・卑語などが多く含まれること，中高教科書で頻出する日常語の多くが落ちていること，初級の読本に対するカバー率が低いことなどが理由である。そのほか，上位 1000 語レベルを補う資料として，数詞，曜日，月名，国名，地名，敬称，略語などを含む 250 語が別途作成されている。

SVL 12000（アルク）は，出版社のアルクが，蓄積してきた様々な英文データと多数の先行資料をもとに，日本人の英語学習者にとって有用であると思われる英語語彙 1 万 2000 語を選び，基礎から上級へと 1000 語区切りの 12 のレベルに区分した段階別学習語彙リストである。語彙の選定は，コーパスの使用頻度をベースにしながら，中学生から一般社会人までの日本人英語学習者にとっての「有用性」「重要性」を考慮して行われている。

コーパス頻度に基づいて作られた語彙表でも，GSL と JACET8000 および SVL12000 の作られ方や目的は異なっている。GSL は，英語において最も役立つ 2000 の基本語彙を選定することを目的とし，多義語の頻度情報までカバーしている。一方，JACET8000 や SVL12000 は，特に日本人の学習者が学ぶべき語彙をある程度網羅的に示したものであり，選定された語彙のレベル分けも行われているが，多義語についてまでは対応していない。いずれも，それぞれの目的に応じて有用な語彙表である。

JACET 8000 タイプの語彙表では，頻度によって語彙が選定され，教育的観点による順位補正が行われることが一般的である。これは，自然言語の頻度が，学習者にとっての難易度をそのまま反映しているわけではないことを示している。教育的観点による順位補正は，対数尤度や散布度のような統計指標を利用したり，英語教科書の出現頻度を加味した調整を加えるなど，語彙表の目的や用途に応じてそれぞれ妥当と考えられる手法で行われている。

また，英語教育の語彙表作成においては BNC が利用されることが少なくない。BNC は 1 億語規模の均衡コーパスである。このことからも分かるように，基本語彙の選定は均衡性の保たれたコーパスに基づいていることも大切である。語彙の使用頻度はジャンルによって特徴があるので，例えば，新聞のみ，小説のみ，などのいわゆる広義のコーパスの頻度をそのまま利用すると頻度順位にもその特徴が反映してしまうことになるからである。

2.2.2.3. 語彙表作成に利用される統計指標

本研究では統計情報を活用し，語彙の客観的選定を行うことを目的としている。コーパスに準拠した語彙表は，コーパスの出現頻度をその基礎資料とする。例えば，英語教育では，1964 年に Brown コーパスが開発されて以来コーパス・データに基づく語彙表の作成が次第に盛んになり，Kučera & Francis (1967)，Hofland & Johasson (1982)，Leech et al. (2001)など，様々なコーパス準拠の語彙表が開発されてきた。また，日本で開発された JACET 8000 や SVL12000 は BNC の頻度リストに基づいている。

このように，コーパスに基づく教育語彙表の作られ方は出現頻度を基礎資料とするが，サブコーパスによって語彙の出現頻度や分布状況は異なるため，ただ出現頻度を集計しただけではテキスト依存する語が高頻度になるなど，サブコーパスの傾向が直接的に影響してしまう。そのため，コーパス準拠の語彙表作成では，通常，散布度

（dispersion）や有用度指標（utility measure）が利用される。

散布度とは，データの分布を表す指標である。散布度の指標は複数あるが，投野・本田（2016）では，よく用いられる指標として以下の(1)～(8)を紹介している。今，あるコーパスが n 個の（または n 分割された）サブコーパスをもつと仮定し，その各サブコーパスが全体に占める割合を$s_1, s_2, \dots, s_i, \dots s_n$（値は百分率）とする。言語特徴 a の各サブコーパスの頻度を$v_1, v_2, \dots v_i \dots v_n$とし，それらの平均を$\bar{v}$，合計の総頻度を$f$とすると：

(1) レンジ（range）：当該言語特徴 a を含む部分の数または割合

(2) 範囲（max-min diff）：$max(v) - min(v)$

(3) 標準偏差（standard deviation）：$sd = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \bar{v})^2}{n-1}}$

(4) 変動係数（variation coefficient）：$vc = {sd}/{\bar{v}}$

(5) カイ 2 乗値（chi-squared）：$\chi^2 = \sum_{i=1}^{n} \frac{(observed\, v_i - expected\, v_i)^2}{expected\, v_i}$

ただし $expected\, v_i = s_i \cdot f$

さらに，コーパス言語学では以下のような指標が古典的なものとして今でも利用されている。：

(6) Juilland's D: $1 - {vc}/{\sqrt{n-1}}$

(7) Carroll's D_2：$\left(\log_2 f - \left(\left(\sum_{i=1}^{n} v_i \log_2 v_i\right)\frac{1}{f}\right)\right)\frac{1}{\log_2 n}$

日本の語彙表では，Tono et al. (2013)が，Carroll's D_2によって散布度と頻度を系統的に示した例がある。

一方, Gries(2008)は既存の散布度指標について以下の五つの問題点を指摘している。第一に，コーパスを構成する部分（corpus parts）のサイズの問題がある。既存の指

標には同じサイズのファイルでなければ計算できないものもある。例えば idf はそのような指標である。しかし，様々なジャンルのテキストから構成されるコーパスのファイルサイズは均一ではないことが多い。第二に，指標が取る値の範囲の問題がある。レンジ，標準偏差，変動係数，カイ 2 乗値は正規化されていないため様々な値を取り，他の研究との比較が難しい。第三に，Juilland's D や Carroll's D_2 のように 0 から 1 の範囲で示される指標とされていても，実際の値はこの範囲を超えることがある。例えば，x という言語特徴が六つのコーパスのうちの一つに 1 回だけ出現する場合，D=－0.095 となる。しかし，Juilland's D の最低値は 0 のはずである。第四に，コーパスを構成する部分の数が散布度の値を左右するが，Juilland's D，Carroll's D2 のように理論的には 0 から 1 の範囲の値を取る指標でも，コーパスを構成する部分の数によっては，全ての部分に同じ頻度で出現する語でも散布度が 1 にならないこともあるし，一つのファイルに 1 回しか出現しない場合でも散布度が 0 にはならないこともある。第五に，指標の感度の問題がある。Juilland's D をはじめとする散布度指標には，実際の語の分布状況が異なるにもかかわらず散布度は同じ値になるなど，必要とする感度を備えていないものもある。

このように，語彙の分布を示す統計指標は過去にも複数開発され利用されてきたが，どれも完璧なものではなく問題点も少なくない。そこで，Gries (2008)では DP という新しい指標を提案している。:

(8) Gries' DP: $$\frac{\sum_{i=1}^{n}\left|\frac{v_i}{f}-\frac{s_i}{\sum s}\right|}{2}$$

DP は，ある言語特徴がサブコーパスに出現する頻度をコーパス全体に出現する総頻度で割った値（実測値）と，そのサブコーパスの総語数をコーパス全体の総語数で割った値（期待値）の差を求め，その絶対値を足し合わせたものを 2 で割るという方法で算出される。DP の示す値の範囲は 0 以上 1 以下で，0 に近いほど分布が安定し，

反対に 1 に近いほど不安定であることを示す。Gries(2008)は，DP を既存の指標の欠点を解消し，かつ，計算方法も容易なものとして提唱している。

次に，有用度については，投野・本田（2016）では，Juilland's usage coefficient U（Juilland, 1964）と Carroll's U_m（Carroll, 1970）を紹介している。有用度指標（utility measure）は，頻度と分布統計を合成して算出するものである。

(9) Juilland's usage coefficient U: $D \cdot f$　　D：(6)を参照

(10) Carroll's U: $(\sum_{i=1}^{n} v_i) \cdot D_2 + (1 - D_2) \cdot \frac{f}{n}$　　D_2：(7)を参照

(10)を利用した語彙表には，Carroll, et al. (1971), Zeno et al. (1995) などがある。

2.2.3. 日本語のコーパス

2.2.3.1. 日本語コーパスの概観

次に日本語のコーパスについて概観する。現在，一般公開されている日本語の電子化されたデータの主なものを表 8 に示す。なお，ここでは大規模で均衡性が保たれた狭義のコーパスだけでなく，言語データベースを広く含む広義のコーパスについても触れる。

表 8　主な日本語のコーパス

コーパス名	年	概要	作成機関／作成者
書き言葉			
新潮文庫の 100 冊	1995	新潮文庫の小説 100 作品を収録	新潮社
新聞記事データベース	1987～	新聞の記事のデータ版	日外アソシエーツ
青空文庫	1997～	Web 上の電子図書館（小説）	青空文庫
国会会議録検索システム	1999～[7]	1947 年以降の国会議事録をウェブから検索可能	国立国会図書館

[7] データベースとしていつから使用可能になったのかは明示されていないが，第 145 回国会（1999 年 1 月開会）以降は会議録原稿をもとに作成されたデータを使用，それ以前の分は発行された会議録を機械で読みとって作成しているため，ここでは 1999～とした。

京都テキストコーパス	1997	毎日新聞の記事に各種言語情報を付与したテキストコーパス。95 年 1 月 1 日から 17 日までの全記事約 2 万文，1 月から 12 月までの社説記事約 2 万文，計約 4 万文に対して形態素・構文情報を付与している。	河原大輔， 黒橋禎夫
現代日本語書き言葉均衡コーパス（BCCWJ）	2011	総語数約 1 億語の大規模現代日本語書き言葉均衡コーパス。新聞，書籍，ウェブデータ，行政白書，韻文など，様々な媒体の現代日本語書き言葉テキストを含んでいる。	国立国語研究所
Web データ			
京都大学格フレーム	2009	Web テキストから自動構築した大規模格フレームで，動詞と共起する格助詞と名詞を検索することができる。約 16 億文の日本語テキストから自動構築し，約 4 万用言から構成されている。	河原大輔， 黒橋禎夫
JpWac	2008	コーパス検索システム Sketch Engine(Kilgarriff et al, 2004)に搭載されていた約 4 億語のコーパス。高頻度の一般的内容語 500 語を選び，それらを組み合わせて 5000～6000 の検索（queries）リストを生成し，Google®が返す上位 10 種の URL ページを自動ダウンロードしてコーパスとしたもの。	Irena Srdanović Erjavec, Tomaž Erjavec, Adam Kilgarriff
JpTenTen	2011	Sketch Engine(Kilgarriff et al, 2004)に搭載されている約 100 億語のコーパス。ツールでデータをクロールし，MeCab と Unidic2 で形態素解析し，短単位と長単位アノテーションが付与されている。	スルダノヴィッチ・イレーナ，スホメル・ヴィット，小木曽智信，キルガリフ・アダム
筑波ウェブコーパス・NINJAL-LWP for TWC	2013～2015	ウェブ上の Html テキストから構築した 11 億語のコーパス。それを検索するための仕組みとして，NINJAL-LagoWord Profiler が導入され，検索環境とセットで提供されている。	筑波大学留学生センター 国立国語研究所 Lago 言語研究所
話し言葉			
女性のことば・職場編	1998	1993 年 9 月～11 月に首都圏で収録された音声資料を文字化したもの。対象は有識の 20 代から 50 代の女性 19 人で，職場でのインフォーマルな場面とフォーマルな場面での自然会話を録音している。	ひつじ書房
男性のことば・職場編	2002	1999 年 10 月～2000 年 12 月に首都圏で収録された音声資料を文字化したもの。対象は有識の 20 代から 50 代の男性各世代 5 人で，異なる職種・職場からなる 19 人の協力者に職場でのインフォーマルな場面とフォーマルな場面での自然会話を録音している。	ひつじ書房

日本語話し言葉コーパス（CSJ）	2006	講演を中心とする自発的な独話のデータ。総数 3302 講演（90%がモノローグ，10%が対話，朗読などの音声）で，662 時間，およそ 752 万語分の音声が収録されている。	国立国語研究所
BTS による多言語話し言葉コーパス	2005～2007	日本語母語話者同士の会話と，日本語母語話者と日本語学習者の会話の文字化資料(CD-ROM)	宇佐美まゆみ監修
学習者コーパス			
インタビュー形式による日本語会話データベース（上村コーパス）	1998	OPI テスターが日本語母語話者(54 人)，非母語話者(56 人）計 120 人に行った 15 分の日本語 OPI の文字化テキストを収録したもの。	上村隆一
KY コーパス	1999	OPI テープを文字化した言語資料。被験者は 90 人で，初級から超級までの中国語，英語，韓国語母語話者を対象とする。	鎌田修 山内博之
日本語学習者による日本語作文と，その母語訳との対訳データベース	2001	日本語学習者による日本語作文と作文執筆者本人による母語訳のデータベース。作文データの総数は 1565 件。	国立国語研究所
日本・韓国・台湾の大学生による日本語意見文データベース	2011	日本語を母語とする大学生（134 人）と日本語を学ぶ大学生（台湾 57 人，韓国 55 人）が日本語で執筆した意見文を収録したデータベース。	伊集院郁子
日本語学習者作文コーパス	2013	日本語学習者による日本語作文をアノテーション済みコーパスとして公開。ウェブインターフェイスによる検索環境も用意されている。作文データの総数は 304 件。	李在鎬
日本語教育のためのタスク別書き言葉コーパス	2014	日本人大学生 30 人と留学生（韓国語母語話者 30 人，中国語母語話者 30 人）による 12 のタスクの書き言葉の資料計 1080 編（母語別各グループ 360 編ずつ）を集めたもの。客観的な評価基準に基づき各タスクの達成の可否を判定し，タスクの達成度により，学習者を上位群・中位群・下位群の 3 グループに分けている。	金澤裕之編
その他のコーパス			
日英新聞記事対応付けデータ	1989～2001	1989 年から 2001 年までの読売新聞と The Daily Yomiuri から自動作成された日英対応付けコーパス。	内山将夫
日英対訳文対応付けデータ	2003	Project Gutenberg ，青空文庫，プロジェクト杉田玄白などの作品について日本語文と英語文との対訳文対応を付けたもの。	内山将夫
太陽コーパス	2005	戦前の日本で広く読まれた博文館の総合雑誌「太陽」の５年分の全文テキストコーパスで，総量は約 700 万語。近代日本語の書き言葉が文語体から口語へと移行した時期の日本語を研究するためのコーパス。	国立国語研究所

日本語の調査・研究にコンピュータが用いられた最初の例は，国立国語研究所（1970）『電子計算機による新聞の語彙調査』である（丸山，2009）。しかし，ここで公表されたのは調査結果だけであり，電子化された日本語のデータを公開して共有し，研究者が利用するというようなことはなかった。

日本語コーパスを研究者が利用するようになったのは，1990 年代になってからである。この頃は『新潮文庫の 100 冊』『朝日新聞記事データベース』『CD-毎日新聞』などがよく利用されていた（丸山，2009）。『新潮文庫の 100 冊』は新潮文庫の小説など 100 冊を収録した電子出版物，『朝日新聞データベース』と『CD-毎日新聞』は新聞記事データである。また，音声認識や形態素解析，機械翻訳に有効利用するために，音声データベースとテキストデータベースの開発も行われた。

1990 年代後半になると，コンピュータとインターネットの普及により，コーパスの構築も進んだ。著作権の消滅した文学作品を収集した『青空文庫』もこの時期に開発，公開された。

話し言葉コーパスも構築された。『女性のことば・職場編』と『男性のことば・職場編』は職場での自然会話を文字化してまとめたものである。学習者コーパスもこの時期から開発されている。『インタビュー形式による日本語会話データベース』（上村コーパス）は，日本語の母語話者（54 人）と非母語話者（56 人）の OPI を収録した音声コーパス，『KY コーパス』は，日本語学習者 90 人分の OPI を文字化したコーパスである。

さらに，自然言語処理の分野では，研究用情報を付与したデータが学界の中で共有されるようになった。『京都大学テキストコーパス』は 1995 年に発行された毎日新聞の記事 4 万文に対して，形態素情報，構文情報を付与したコーパスである。これにはその後も様々な情報が付与され現在も利用されている。

2000 年以降に開発されたコーパスの主なものには，『日本語話し言葉コーパス』（Corpus of Spontaneous Japanese：以下，CSJ），『BTS による多言語話し言葉コーパス』，『現代日本語書き言葉均衡コーパス』（Balanced Corpus of Contemporary

Written Japanese: BCCWJ）などがある。

CSJ は，国立国語研究所，情報通信研究機構，東京工業大学により共同開発された大規模な自発音声コーパスである。独話を中心とした日本語の自発音声約 661 時間，752 万語が収録されており，発話の転記テキスト，形態論情報，イントネーションラベル，係り受け構造情報など，様々な研究用情報が付与されている。

『BTS による多言語話し言葉コーパス』は，東京外国語大学大学院地域文化研究科 21 世紀 COE プログラム「言語運用を基盤とする言語情報拠点」により作成・公開された，談話研究を目的とした小規模なコーパスである。日本語母語話者同士の会話と，日本語母語話者と日本語学習者の会話を文字化したもので，同時発話，挿入，相槌，笑いなどが記号によって表現されている。

BCCWJ は国立国語研究所により構築されたコーパスである。言語研究用に設計されていて，1976 年から 2005 年の 30 年間にわたる書き言葉を対象としている。サブコーパスは，書籍，新聞，雑誌，白書，国会会議録，ウェブ上のテキスト，教科書などで，その中の一部はジャンルの構成比率が決められている。

また，Yahoo!や Google などの検索エンジンを使い，インターネット上のウェブページに存在する言語表現を集計し，ウェブページの集合をコーパスとして見なす研究も出てきた。これについては安定性，信頼性の点から言語資料として使うことには懐疑的な見方もあるが，大量の言語データをコーパス化することができるという長所を備え，今後もますます増えていくと予想されている。このタイプのコーパスには，『国会会議録検索システム』，『JpWac』，『JpTenTen』，『筑波ウェブコーパス』，『京都大学格フレーム』などがある。『国会会議録検索システム』は，1947 年の第一回会議以降全ての本会議，委員会などにおける会議録を検索できるサービスで，全体で 35 億文字の規模を持つ。『JpWac』はウェブ上の言語から収集した約 4 億語のコーパス，『JpTenTen』は約 100 億語という巨大な規模のコーパスで，コーパス検索システム『Sketch Engine』（Kilgarriff et al, 2004）に搭載されている。『筑波ウェブコーパス』は，ウェブ上のテキストから構築した 11 億語のコーパスで，NINJAL-LagoWord

Profiler という検索システムとセットで公開されている。また，自然言語処理の分野でも，ウェブをコーパスと見たて，数々の言語情報を得ようとする動きが進んでいる。『京都大学格フレーム』は，ウェブから収集した約 16 億文のテキストを構文解析し，用言と名詞との間で結ばれる格関係を整理したものである。

2000 年以降は学習者コーパスの整備も進んだ。『日本語学習者による日本語作文と，その母語訳との対訳データベース』は 2001 年に国立国語研究所によって作成された，日本語学習者による日本語作文と母語訳のデータベースで，1565 件の作文データを収めている。『日本・韓国・台湾の大学生による日本語意見文データベース』は，日本語を母語とする大学生と日本語を学ぶ大学生（台湾，韓国）が日本語で執筆した意見文を収録したデータベースである。『日本語学習者作文コーパス』は，日本語学習者による 304 件の日本語作文をアノテーション済みコーパスとして公開したもので，ウェブインターフェイスによる検索環境も用意されている。『日本語教育のためのタスク別書き言葉コーパス』は，日本人大学生と留学生（韓国，中国）による 12 のタスクの書き言葉の資料 1080 件を集めたものである。

そのほか，特殊な研究目的のために作られたコーパスもある。『太陽コーパス』は明治から大正期にかけて刊行された雑誌データを集めた通時コーパスである。また，日英のパラレルコーパスには，『日英新聞記事対応付けデータ』や『日英対訳文対応付けデータ』などがある。

このように，日本語コーパスの開発と利用は 1990 年代から始まり，2000 年以降盛んになっている。今後は，BCCWJ のように言語研究用に設計された大規模コーパスの開発および分析と，ウェブ上のテキストを対象とする大規模なテキスト処理技術の開発という二方向に進んでいくと見られている（丸山，2009）。

2.2.3.2. 『現代日本語書き言葉均衡コーパス』（BCCWJ）について

このように様々なコーパスが作成されてきたが，日本語の書き言葉コーパスの代表

とも言えるのが BCCWJ である（李，砂川，2012)。これは，国立国語研究所の言語データベース整備計画である KOTONOHA 計画[8]の一環として現代日本語の書き言葉を正確に代表する均衡コーパスで，言語研究における利用を目的としている。コーパスの代表性（representativeness）を確保するため，BCCWJ では新聞または小説など一つのジャンルに限定せず，広範囲の様々に異なるテキストを収集し，現代日本語書き言葉全体の「縮図」となるようなコーパスを作成することを意図している。

BCCWJ の内部構成は以下の通りである。BCCWJ のサブコーパスは，出版（生産実態）サブコーパス，図書館（流通実態）サブコーパス，特定目的（非母集団）サブコーパスの三つに大別される。出版サブコーパスは，2001 年から 2005 年までに出版された書籍，雑誌，新聞の文字の総体を母集団とし，そこから約 3500 万語相当のテキストを無作為抽出したサブコーパスである。図書館サブコーパスは，東京都下 52 自治体の公立図書館（1 自治体＝1 図書館と見なす）のうち 13 館以上に所蔵されている書籍で，ISBN が付与されており，1986 年から 2005 年の期間に出版されたものの全体を母集団とし，約 3000 万語のテキストを無作為抽出したものである。図書館コーパスは，単に出版されただけでなく，ある程度まで広い範囲に流通したことが確実である書籍を母集団としていること（したがって，特殊な専門書や公序良俗に反する内容の書籍が排除されていること)，および，対象期間が 20 年に渡ることの 2 点において，出版サブコーパスの書籍部分と異なっている。特定目的サブコーパスは，日本語研究上大切ではあるが，出版ないし図書館サブコーパスのサンプリングによっては十分な数が集まらないと考えられるものや，そもそも母集団を定義してサンプリングすることが不可能なものを集めたものである。ここには，「白書」「Yahoo!知恵袋」「国会会議録」「ベストセラー」「検定教科書」「ブログ」のデータがある。「白書」は過去 30 年間に政府が刊行した白書，「Yahoo!知恵袋」はインターネット掲示板のデータから，「国会会議録」は国会図書館がインターネットで公開している衆参両院の議事録のうち過去 30 年分を対象としたもの，「ベストセラー」は過去 30 年間にベストセラー

[8] KOTONOHA 計画とは，国立国語研究所の日本語データベースの長期整備計画で，明治から現代に至るまでの近現代日本語が対象となっている。

リストに載った書籍 951 冊を対象にしたもの，「検定教科書」は小中高の検定教科書，「ブログ」は Yahoo!ブログのうち書き込みが多いもの等の条件を満たしたブログのテキストから，いずれも無作為抽出されたテキストである。

表 9　BCCWJ の構成

出版（生産実態）サブコーパス 出版された書籍，新聞，雑誌 対象期間：2001-2005 年 3500 万語	図書館（流通実態）サブコーパス 東京都の 13 自治体以上の図書館に所蔵されている書籍 対象期間：1986 年-2005 年 3000 万語
特定目的（非母集団）サブコーパス Yahoo!知恵袋，Yahoo!ブログ，白書（500 万語），国会会議録（500 万語），ベストセラー，検定教科書 対象期間は様々。最長 30 年間。 3500 万語	

（前川・山崎，2009 の表を編集した）

表 10　BCCWJ 出版サブコーパスの内訳

	構成比	サンプル数	語数（可変長9）
書籍	74.1%	12604	2891.5 万語
雑誌	16.1%	2730	481.8 万語
新聞	9.8%	1666	98.0 万語
合計	100.0%	17000	3471.3 万語

（前川・山崎，2009 の表を編集した）

なお，BCCWJ は最長で過去 30 年間の日本語を対象資料としているが，これが現代日本語の範囲として最も適当であると判断できる科学的根拠があるわけではない。これは，母集団設定の容易さ，データ入手の容易さなどから検討して得られた経験値によって決められたものである（前川・山崎，2009）。しかし，BCCWJ が従来にはなかった日本語書き言葉均衡コーパスであり，日本語研究や日本語教育研究において利用価値の高いものであることは間違いない。

9 BCCWJ のサンプル長は，一律 1000 語に固定した「固定長サンプル」と，節や章など文章構成上の単位と対応し，意味的なまとまりを持った「可変長サンプル」がある。

2.2.4. コーパスに基づく日本語教育語彙表

コーパスに基づく日本語教育語彙表はまだあまり多くない。英語では GSL のようなリストが 1950 年代から利用可能であったほか，BNC のような大規模均衡コーパスに準拠した語彙表も複数作られてきた。しかし，日本語の場合 90 年代に入るまでコーパスが簡単に利用できなかったため，コーパス準拠の教育語彙表を作る研究は比較的新しいものである。コーパスを利用した日本語教育語彙表には表 11 のようなものがある。

表 11　コーパスを利用した主な日本語教育語彙表

語彙表名	年	概要	総語数	作成者
日本語能力試験（新試験）出題基準語彙表 （非公開）	2010	2010 年に改訂された日本語能力試験に合わせ，試験問題作成者用に作られた語彙表。コーパスや語彙調査等の資料を利用し，専門家判定方式によって語彙を選定。語彙表そのものは非公開だが，語彙表作成に関する論文が発表されている。	約 17000	国際交流基金
日本語を読むための語彙データベース	2011	BCCWJ モニター公開データ（2009 年度版）の書籍および「Yahoo 知恵袋」（約 3300 万語）で作成された語彙リスト。コーパス頻度およびサブコーパスごとの語彙分布をもとにして，語彙を「一般用」，「留学生用」に分け，レベル分けをしている。	60894	松下達彦
日本語教育語彙表	2012	『現代書き言葉均衡コーパス』（BCCWJ）モニター公開データ（2009 年度版）と「日本語教科書コーパス」（初級から上級まで市販されている教科書 100 冊の電子データ版；非公開資料）を語彙表作成における基礎として使用。日本語教育上の習熟度，語彙レベルが入っているほか，頻度情報をもとに，5 段階で重要度を示している。	約 18000	砂川有里子 李在鎬
A Frequency Dictionary of Japanese (Routledge Frequency Dictionaries)	2013	『現代書き言葉均衡コーパス』（BCCWJ）と『日本語話し言葉コーパス』（CSJ）の長単位頻度リストに基づき，5000 の高頻度語を示した頻度辞書。和英辞書として訳語や例文が付いているほか，頻度や統計情報も記載されている。	5000	投野由紀夫 前川喜久雄 山崎誠

まず，語彙表は BCCWJ の公開前に作られたものと，公開後に作られたものに分けて考えることができる。公開前に作られたものには，国際交流基金の「日本語能力試験（新試験）出題基準語彙表」（秋元・押尾，2008，押尾・秋元・武田・阿部・高梨・柳沢・岩元・石毛，2008）がある。公開後に BCCWJ を利用して作られた語彙表に，「日本語を読むための語彙データベース」（松下，2011），「日本語教育語彙表」（李・砂川，2012），*"A Frequency Dictionary of Japanese"* (Routledge Frequency Dictionaries) (Tono et al.,2013)がある。

秋元・押尾（2008）および押尾他（2008）は，2010 年に改訂された日本語能力試験に合わせ，試験問題作成者用に作られた語彙表である。コーパスや語彙調査等の資料を利用しつつも，客観的な語彙選定というよりは，専門家判定方式の特色が強いものと考えられる。語彙表そのものは，一般公開されていないため，どのような内容であるかは明らかではないが，押尾他（2008）は，新試験の語彙表作成の方針として以下の 3 点を挙げている。:

(1) 客観的かつ大規模な語彙データベースを複数組み合わせた資料から語を選別し，初級～中上級までの一覧表を作成する。
(2) 書き言葉だけでなく話し言葉もこれまで以上に考慮する。
(3) 描写を豊かにする様々な表現も積極的に採用する。

また，選別の方針として以下の 4 点を挙げている。:

(1) 主に頻度を重視して採否を決める
(2) 機械的に頻度の高いものから採用するのではなく，日本語教育経験者の視点も加える。
(3) 現行試験の「出題基準」語彙表も参考にする。
(4) 最終的な語数は，日本人成人の獲得語数などを参考にした上で決定する。

このような方針に基づき，「大規模な語彙データベース」（表 12）が作成され，そこから語彙の選定が行われている。

表 12　国際交流基金が新試験対応語彙表作成のために使用したデータベース

				発行年	出版
第三次DB	第二次DB	第一次DB	『日本語の語彙特性第１期』CD-ROM	1999	三省堂
			『現代新国語辞典改訂版第３版』CD-ROM	2000	学研
			『日本語の語彙特性第２期』CD-ROM	1999	三省堂
			『出題基準（改訂版)』	2004	凡人社
			『現代雑誌の語彙調査 1994 年発行 70 誌』	2005	国立国語研究所
		外来語	『日本語の語彙特性第１期』CD-ROM	1999	三省堂
			『例解新国語辞典』	1999	三省堂
			『現代新国語辞典改訂第３版』	2002	学研
			外来語認知に関する調査	2002～2004	国立国語研究所
			『出題基準（改訂版)』	2004	凡人社
			『現代国語例解辞典』	2006	小学館
		オノマトペ	『擬音語・擬態語辞典』	1974	東京堂
			『擬音語・擬態語辞典』	1978	角川書店
			『絵でわかる　ぎおんご　ぎたいご』	1994	アルク
			『現代擬音語擬態語用法辞典』	1978	角川書店
			『現代国語例解辞典』第４版　擬音語・擬態語集成	2005	小学館
	話し言葉		「男はつらいよ」全 48 編スクリプト	1969～1995	CASTEL/J 研究会
			話し言葉資料	1997～2001，2004	筑波大学
			『女性の言葉　職場編』	1999	ひつじ書房
			『男性の言葉　職場編』	2002	ひつじ書房
			BTS による多言語話し言葉コーパス―日本語会話（１）（２）（2003 年版）	2003	東京外国語大学宇佐美まゆみ監修
慣用表現など			『小学国語学習辞典』	1994	偕成社
			『小学国語新辞典』第３版	2002	旺文社
			『くもんの学習国語辞典』第３版	2002	くもん出版
			『例解新国語辞典』第６版	2002	三省堂
			『現代新国語辞典』改訂第３版	2002	学研
			『例解学習国語辞典』第８版ワイド版	2004	小学館
			『例解小学国語辞典』第３版	2005	小学館

（押尾他，2008 をもとに編集した）

ここで作られた「大規模な語彙データベース」はコーパスだけではなく，多くは，書き言葉，外来語，オノマトペ，話し言葉という枠組みから選ばれた辞典や既存の語彙表などである。これらからは語彙の頻度情報は得られない。したがって，「(1)主に

頻度を重視して採否を決める」とあるが，その頻度情報は『日本語の語彙特性第 2 期』，『現代雑誌の語彙調査 1994 年発行 70 誌』などから得られたもの，すなわち，新聞や雑誌という限定されたジャンルのコーパスの頻度情報である。

語の選別作業は，次の三段階で行われている。まず，第一段階では，第一次 DB（約 12 万語）から，『日本語の語彙特性第 2 期』（朝日新聞の 14 年分（1985 年～1998 年）の記事データ）の出現頻度，『現代雑誌の語彙調査 1994 年発行 70 誌』の出現頻度，単語親密度，「出題基準」の初出級，『現代新国語辞典』の採録情報などに基づき，「学習者が知っていたほうがよいかどうかという選別基準」で，選定者が一語一語採否を決め，約 1 万 7 千語が選別されている。

第二段階では，外来語，オノマトペ[10]，日本語教育的観点による必要性から選ばれた語彙を語彙データベースに加え（第二次 DB），「選別基準に沿った再検討」が行われている。「選別基準」は全て公開されてはいないが，秋元・押尾（2008）では，「国名や地名は載せず，大陸名や地域名に限って載せる」や，「日，月，年の言い方は，基点の前後二つまでを載せる」などの形式的な基準が一例として示されている。

第三段階では，話し言葉データベースが加えられ，「レベルイメージ」をもとにレベル分けがされている。「レベルイメージ」とは，新試験の N1，N2，N3，N4，N5 の 5 段階のレベルに関する can-do statements のイメージのことである。つまり，選別者が語彙を can-do statements のイメージに相応しいと思うレベルに選定していくという方法で，最終的な語彙レベルの設定が行われている。

したがって，新試験の語彙リストの語彙は，部分的にコーパスを利用しつつも，頻度情報や統計を主要な基準として選ばれたものではなく，基本的には旧試験の「出題基準」のように主に選定者の主観や直感によって選ばれ，レベル分けされた専門家判定方式タイプのものと考えられる。作成当時，BCCWJ のような大規模均衡コーパスが利用できなかったため，単語親密度や can-do statements のイメージを利用するな

10 オノマトペとは，「外界で発せられる声と音を移した言葉である擬音語と，ある動きや状態などを音によって抽象的に表す言葉である擬態語，また，擬情語と呼ばれる人の心の状態を表す言葉を総称した語」である（押尾他，2008）。

どの日本語教育的観点からの様々な工夫が施されている。

次に，BCCWJ を利用したものには，松下（2011），李・砂川（2012），Tono et al. (2013) がある。

松下（2011）は，BCCWJ モニター公開データ（2009 年度版）の書籍（BK）および「Yahoo!知恵袋」（OC）（約 3300 万語）で作成された語彙リストである。語の単位は Unidic の短単位が使われている。コーパス頻度およびサブコーパスごとの語彙分布をもとにして，「書きことば」の重要度ランキングとレベル分けを行っている。これは，具体的には，100 万語あたりの使用頻度に散布度[11]を掛けたものによる。さらに，語彙を「一般用」，「留学生用」に分け，初級から超上級までのレベル分けをしている。「留学生用」とは，日本の大学で学ぶ留学生を対象として想定したものである。一方，「一般」はより幅広い一般的な目的で日本語を学ぶ学習者の利用を想定したものである。「留学生用」も「一般用」も，「書きことば重要度ランク」をベースに，旧日本語能力試験「出題基準」語彙表の語彙とそのレベル分けを利用して，ランク付けとレベル分けが行われている。「留学生用」のほうは，「初級のみ日常生活を意識し，それ以外は書きことばを中心に」考えられている。すなわち，初級以外の部分は「書きことば重要度ランク」の順に配列されている。また，「一般用」のほうは，「全レベルにわたって日常生活を意識して考えた語彙」としている。

松下（2011）は，日本の留学生向けとその他一般の日本語学習者に分けて語彙の難易度設定をし，6 万語以上の多数の語彙に対し統計指標によって定量化した重要度を提示している点で，日本語教育においては新しい試みと言える。しかし，そこで使われているコーパスは BCCWJ のうちの書籍（BK）と Yahoo!知恵袋（OC）だけであり，日本語教育語彙表を作成する元となるコーパスとして適当な語彙を抽出できるかという点については検討されていない。松下（2011）で使われたコーパスは，書籍と Yahoo!知恵袋の日本語を読む上での語彙の「重要度ランク」を提示できる材料ではあるが，その他の多様なテキスト媒体についてまで同様のことを示せるものではない。また，

11 Juilland's D を使用している。

「留学生用」「一般用」など，特に日本語教育を意識した語彙の選定やレベル設定については「出題基準」を参照している部分もある。したがって，松下（2011）は，コーパスの出現頻度と分布統計から客観的に語彙の重要度を提示しつつも，語彙のレベル設定では過去の専門家判定方式を部分的に引き継いだ形である。

李・砂川（2012）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」（代表者：砂川 有里子，2011 年度～2014 年度）において，学習者向け辞書開発の基礎資料として開発されたもので，リストの総合数は約 1 万 8 千語である。コーパスは，独自に開発された日本語教科書 100 冊の「日本語教科書コーパス」および BCCWJ 2009 年度版領域内公開データを利用している。語の単位は Unidic の短単位が使われている。見出し語を決定は，形態素解析後，助詞類や助動詞類を取り除き内容語のみにしたうえで，出現頻度 5 以上のものをリスト化して行っている。さらに，見出し語では，形態素解析（Unidic-MeCab）による語彙素に加え，形態素 N-gram[2] を使用し，形態素を結合させる方法で一部の複合語も抽出している。語のレベル分けでは，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という主観的方法がとられ，語彙の難易度は，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 段階に分かれている。語彙表には，見出し語，品詞，難易度レベル，語種などの情報のほか，分類語彙表による意味分類，「出題基準」の級，コロケーション情報，語義など，様々な情報が付与されている。

李・砂川（2012）は，コーパス出現頻度と専門家判定方式で語彙の選定とレベル設定を行っているが，松下（2011）と同様に，コーパス自体が語彙表の目的に合うものかどうかを検討したという記述はない。また，語彙の難易度は機械的に決められるものではなく，日本語教師の「経験」や「勘」が反映される必要がある（李，2013, p.13）という立場から，従来通りの専門家判定方式が採用されている。

Tono et al. (2013) は，BCCWJ と CSJ の長単位頻度リストに基づき，5000 の高頻度語を選定した頻度辞書である。和英辞書として訳語や例文が付いているほか，頻度や統計情報も記載されている。コーパス準拠の語彙表は短単位で集計を行っているも

のが多いなか，この語彙リストは長単位を採用しており，合成語や複合辞を抽出している点が特徴的である。これは日本語教育的観点から見ると非常に有用で，他の語彙表では得られない合成語の重要性や複合辞の頻度情報なども確認することができる。また，散布度[12]が付与され，語彙の分布状況が参照できるほか，分野語がコラムとしてまとめられるなどの教育的工夫が施されている。しかし，Tono et al.（2013）は頻度辞書であり，頻度順に上位 5000 語を提示しているものなので，日本語教育的観点からの重要度によってリストを補正するというような作業は行われていない。

このように，コーパスに基づく日本語教育語彙表でもコーパス自体の検討がなされている例はない。しかし，教育のための語彙表作成におけるコーパス利用では，「入れたものが出てくる」というコーパスの必然的な特徴を十分理解して，コーパスに含める資料の内容の吟味は適切に行われなければならない（投野・本田，2016）。日本語教育基本語彙の選定を目的とするならば，その目的に応じてコーパス自体の中身の検討が十分に行われることは非常に重要である。

語彙の選定とレベル設定では，秋元・押尾（2008），押尾他（2008），松下（2011），李・砂川（2012）いずれの例も，コーパスの頻度情報を利用していても専門家判定方式を使っているか，過去に専門家判定方式で作られた語彙表の基準を参照している。このことは，日本語教育において専門家判定方式が高く評価され信頼されていることを示唆している。だが，2.1.4.でも述べたように，専門家判定方式で選ばれた複数の基本語彙の間の一致率は高くない。従来からこの方法で数多くの語彙表が作られ，日本語教育研究では成果を残したが，語彙選定における客観性の点では疑問が残る。そして，コーパス準拠であることの利点は，語彙選定とレベル設定を客観的に行えることにある。現在では，大規模コーパスの客観的頻度データと，かつては主観的に行ってきた語彙の精選，加除の作業をどう合理的に組み合わせてゆくかが語彙表開発の課題となっている（石川，2008）という指摘もある。

12 Carroll's D_2が採用されている

2.3節　日本語コーパス研究

2.3.では，日本語コーパス研究から得られる日本語教育語彙表作成の指針についてまとめる。日本語コーパス研究は幅広いが，ここでは語彙表作成と直接関わりのある語の単位，表記，文体ジャンルに絞って述べる。

2.3.1.では，語の単位について述べる。日本語には分かち書きの習慣がなく，どのような単位を一語と見なすかの判断は容易ではない。そのため，従来から語彙調査やコーパス構築において有用な単位が複数開発されてきた。しかし，これらは日本語教育を目的として考案されたものではないため，いずれも日本語教育で一般的に用いられている語の単位とは異なる部分がある。

2.3.2.では，表記の問題について述べる。語の単位の問題と同様，自動解析によって出力される表記も，機械処理上の合理性に基づいて設定されたものであり，教育面を考慮して作られたものではない。そのため日本語教育語彙表作成にあたっては，自動解析された結果の表記を日本語教育に適した表記に補正する作業も必要である。ここでは自動解析システムによって出力される語の表記と日本語教育で一般的に用いられる語の表記を比較し，日本語教育語彙表の見出し語との違いについてまとめる。

2.3.1. 語の単位

2.3.1.1. 日本語コーパスの語の単位

語をどのように定義するかには様々な立場がある。そのため，コーパスの言語単位をどのように規定するかについても様々な立場がある（小椋他，2007)。特に，日本語には分かち書きの習慣がないため，どのような長さを一語と見なすかの判断が難しい。このような背景から，過去の語彙調査では，短単位（β単位)，長単位，M 単位，W 単位などの調査単位が開発されてきた。以下，それぞれがどのような単位であるか，例を示す。:

(1) 短単位（β単位）　型紙/どおり/に/裁断/し/て/外出/着/を/作り/まし/た/。

(2) 長単位　型紙どおり/に/裁断し/て/外出着/を/作りました/。

(3) M[13]単位　型/紙/どおり/に/裁断/し/て/外出/着/を/作り/まし/た/。

(4) W[14]単位　型紙どおり/に/裁断して/外出着/を/作りました/。

（国立国語研究所，1984，p.81）

これらは大まかに形態素相当の短い系列の単位と文節相当の長い系列の単位に分けられる。現在ではこのうち短単位と長単位がコーパスに利用されており，CSJ とBCCWJ には短単位と長単位の情報が付与されている。

短単位は『現代雑誌九十種調査』の β 単位を，長単位は『テレビ放送の語彙調査』の長い単位をもとに設計されたものである（小椋他，2007)。短単位と長単位はそれぞれの研究目的によって使い分けられている。例えば，短単位は用例収集や基本語彙の選定に適した単位であり，長単位は媒体別・分野別の言語特徴を捉えるのに適している（小椋他，2007)。これまで語の調査単位が複数開発されてきたが，これらの単位が言語単位としてどういう位置づけなのかは明らかにされていない。各語彙調査での単位認定の基準を参照しても,「このような条件のものを調査単位とする」という操作的な定義が書いてあるだけで，その言語学的な意味づけは何かということは説明されていない。

次に，短単位と長単位の認定規定について詳しく見ていく。短単位は言語の形態的側面に着目して規定された単位である。したがって，短単位では，語はまず意味を持つ最小単位に分けられる。意味を持つ最小単位は，和語・漢語・外来語・記号・人名・地名の種類ごとに次のように規定されている。最小単位とは,次のような単位である。なお,「/」は最小単位の境界,「｜」は短単位の境界,「┃」は長単位の境界,「=」は切らないことを示す。:

13 形態素 morpheme の略

14 語：word の略

(1) 和語：/豊か/な/暮らし/に/つい/て　　/大/雨/が/降っ/た/の/で/

(2) 漢語：/国/語/　/研/究/所/

(3) 外来語：/コール/センター/　　/オレンジ/色/

(4) 人名：/星野/仙一/　　/ジェフ/・/ウィリアムス/　　/林/威助/

(5) 地名：/大阪/府/豊中/市/待兼山町/　　/六甲/山/　　/琵琶/湖/

(6) 記号：/図/A/　/JR/

（小椋他，2007, p.103）

小椋他（2007）では，短単位の認定規定を以下のように説明している。短単位認定において上記の最小単位は表 13 のように分類されている。

表 13　最小単位の分類

分　類		例
一　般		和　語：豊か　大　雨 漢　語：国　語　研　究　所 外来語：コール　センター　オレンジ…
数		一，二，十，百，千…
その他	付　属　要　素	接頭的要素：相　御　各…
	助詞・助動詞	接尾的要素：兼ねる　がたい　的…
	人名・地名	う　だ　ます　か　から　て　の
	記号	A　B　ω　イ　ロ　ア　JR

（小椋他，2007，p.103）

短単位はこの最小単位を次のような認定規定（一般，数，その他）に基づいて結合させることにより認定されている)。:

[1] 一般

《原則》

(1) 和語・漢語は，2 最小単位の 1 次結合体を 1 短単位とする。

|母=親|　|食べ=歩く|　|言=語|資=源|　|研=究|所|　|本=箱|作り|

⑵ 外来語は，1最小単位を1短単位とする。

|コール|センター|　|オレンジ|色|

《例外規定》

⑴ 省略された外来語の最小単位の扱い

① 省略された外来語の最小単位は，和語・漢語の最小単位と同様に扱う。

|パソ=コン|　|塩=ビ|　|ピン=ぼけ|

② 省略された外来語の最小単位と省略されていない外来語の最小単位との1次結合体は1短単位とする。

|エア=コン|　|マス=コミ|

⑵ 1最小単位を1短単位とするもの

① 最小単位の3個以上の結合体を1短単位とするもの

|衣|食|住|　|松|竹|梅|　|都|道|府|県|

② 類概念を表す部分と名を表す部分とが結合してできた固有名詞のうち，類概念を表す部分と名を表す部分とが共に1最小単位の場合の，それぞれの最小単位。

|さくら|屋|　|歌舞伎|座|　|のぞみ|号|

[2] 数

「数」以外の最小単位と結合させない。「数」どうしの結合は，一・十・百・千のとなえを取る桁ごとに 1 短単位とする。「万」「億」「兆」などの最小単位は，それだけで1短単位とする。小数部分は1最小単位を1短単位とする。

|十二|月|二十三|日|　|七百|五十|二万|語|　|五|分|の|二|　|二三十|回|　|〇|.|四|五|

[3] その他

1最小単位を短単位とする。

付属要素：|筒|状|　|扱い|兼ねる|

助詞・助動詞：|豊か|な|暮らし|に|つい|て|

人名：|星野|仙一|　　|ジェフ|・|ウィリアムス|　　|林|威助|

地名：|大阪|府|豊中|市|待兼山町|　　|六甲|山|　　|琵琶|湖|

記号：|図|A|　　|JR|

これに対し，長単位はまず文節に分割し，それによって得られたものを 1 単位とするような単位である。長単位の認定規定については，小椋（2006）に以下のように記されている。:

[1]　付属語

⑴ 付属語は 1 長単位とする。

┃今┃は┃ファックス┃とか┃そう┃いう┃の┃が┃ある┃んです┃けれども┃

⑵ 形容動詞及び形容動詞活用型の助動詞(そうだ・みたいだ・ようだ）の活用語尾は助動詞として扱い，1 長単位とする。

┃統一的┃な┃視点┃で┃切り┃ましょ┃う┃

┃涙┃が┃出┃そう┃に┃なる┃

┃エンジニア┃な┃んだ┃そう┃です┃

┃駅員さん┃が┃いる┃みたい┃だ┃

┃使える┃よう┃に┃し┃たい┃

⑶ 文節の認定の際に一続きとして扱うこととした固有名・動植物名・「―の～」「―が～」の体言句・分数の読み上げの内部にある助詞・助動詞は切り出さない。

┃西=が=丘┃┃サキシマスオウ=ノ=キ┃┃絵=の=具┃┃万=が=一┃

┃三分=の=二┃┃後続単語種類数分=の=先行単語頻度┃

[2] 並列及び同格の関係にある語は互いに切り離す。

┃安心┃確実┃な┃方法┃

┃塩┃こしょう┃を┃かける┃

┃機関誌┃計量国語学┃

並列及び同格の関係にある体言連続のうち，並列された体言全体に係る体言接辞がある場合は切らない。また並列された体言全体を受ける体言・接辞・形式的な意味の「する」「できる」「なさる」「いたす」がある場合は切らない。

┃平成=九年=十年┃ ┃関東=東北=地方┃ ┃機関誌=計量国語学=発行┃

┃観察=整理=する┃

[3] 体言（合成語）の一部分が連体修飾語を受けている場合，その部分の後で切る。

┃項構造┃の┃曖昧性┃解消┃

「以降」「間(かん)」「ごと」「自体」「達」が付いた場合は切らない。

┃文章┃の┃途中=以降┃ ┃住ん┃でる┃人=達┃

[4] 体言及び副詞に形式的な意味の「いたす」「する」「できる」「なさる」が直接続く場合，体言及び副詞と「いたす」「する」「できる」「なさる」との間は切らない。

┃許容=する┃ ┃演出=できる┃ ┃体験=なさる┃ ┃きらきら=する┃

┃きちんと=する┃

ただし，前の体言が連体修飾を受けている場合は用言部分を切り離す。

┃面白い┃説明┃する┃人┃

[5] 「お(ご) ＋動詞連用形(名詞) ＋する・くださる・いただく・なさる・いたす・ねがう・もうしあげる・あそばす」は全体で一続きとする。

┃お=会い=する┃ ┃お=与え=ください┃ ┃お=電話=なさる┃

┃御=登場=願う┃

[6] 数量を表す要素を含む自立語は，以下のように処理する。

(1) 前の要素に関する順序・番号を直後の要素が表している場合，両者を切り離さない。

┃昭和十三年=八月=八日┃ ┃朝=八時┃ ┃予稿集=八十七ページ┃

┃入所=二十年目┃

(2) 上記の規則に該当しない場合，数量を表す要素とその直前の要素とを切り離す。

┃果汁┃百パーセント┃　　┃バニラエッセンス┃少々┃

┃山の手線┃京浜東北線┃二本┃　┃一箱┃三万┃　┃週┃二通┃

┃一学年┃上┃　┃十年以上┃前┃　┃延べ┃百二十九文┃´

ただし，数量を表す要素が前で列挙された要素の個数を表しているものについては，数量を表す要素と前の要素とを受ける体言がある場合，切り離さない。

┃果汁=百パーセント=オレンジジュース┃

以上，日本語におけるコーパスの語の単位について，短単位と長単位を中心に見てきた。これらは機械処理上の合理性に基づいて設定されたものであり，教育面を考慮して作られたものではない。そのため，日本語教育にこれを応用する場合には，短単位や長単位の性質を理解し，必要に応じて修正すべきである。

2.3.1.2. 日本語教育の語の単位とコーパスの語の単位

コーパス日本語学では短単位と長単位が主な語の単位として利用され，特に短単位は基本語彙の選定に適した単位とされている。しかし，これらはそもそも日本語教育を目的として作られたものではないため,日本語教育で使うには不都合な部分もある。以下，具体例を挙げて説明する。

短単位ではその性質上，合成語をほとんど抽出することができない。例えば,「お母さん」は日本語教育では通常 1 語として扱われるが，短単位では｜お｜母｜さん｜に 3 分割され,「お母さん」という語の抽出ができない。｜積極｜的｜｜図書｜館｜｜駐車｜場｜｜卒業｜式｜｜日本｜語｜｜月曜｜日｜なども同様である。いずれも日本語教育では通常 1 語として扱われるものであるが，短単位ではこのような形で抽出できず分割される。

また，短単位から頻度合計を出すと上記の例ように合成語も分割されるため，日本語教育的には，各語の頻度を直接利用できない部分もある。例えば,「ついて」という複合辞[15]を短単位にすると，動詞「就く」と助詞「て」に分割される。しかし，このために動詞「就く」の頻度が高くなったとしても，日本語学習者にとって動詞「就く」を学習する重要度がそれだけ高いということにはならない。また，例えば,「消極的」は「消極」と「的」に分かれる。「消極」は通常，単独で使用されるのではなく,「消極的」や「消極性」など，接辞とともに派生語として使われることが多い。しかし，短単位で頻度集計を行うと，これらが分割されて集計されるので「消極」という実際には単独であまり使われない語の頻度が高くなり,「消極的」「消極性」はリストに含まれない結果となる。このような例が短単位には多数存在する。

一方，日本語教育的観点から見た場合の長単位の問題は，短単位とは反対に 1 語が長すぎる点にある。例えば，┃御=登場=願う┃や┃駅員さん┃や┃予稿集=八十七ページ┃のようなまとまりが 1 語として扱われる。このように分けると,「登場」と「ご登場願う」,「駅員」と「駅員さん」,「予稿集八十七ページ」と「予稿集」が別語として立てられることになるため，日本語教育語彙表の作成には不向きである。頻度集計の段階でも，これらは見出し語が別であるため頻度も分散してしまう。

どこまでを 1 語としてみなすかという判断は難しい。日本語教育で利用する場合は，目的によってこれらを使い分け，必要に応じて補正しなければならない。

2.3.2. 漢字の表記

コーパスのテキストは,短単位や長単位で単位認定され,形態論情報が付与される。例えば，BCCWJ に付与されている形態論情報には，語彙素，語彙素読み，語種，品詞，活用の種類。活用形などがある。また，CSJ には，代表形，代表表記，品詞，活用の種類，活用形などの情報が付けられている。語彙素や代表表記は漢字等で表され

[15] いくつかの語が複合して，ひとまとまりの形で辞的な機能を果たすもの（松木 1990）

たもので，国語辞典の見出しと見出し語に与えられた漢字表記に相当する。日本語教育語彙リストを作成する場合，その見出し語には形態論情報の語彙素や代表表記を利用することになるが，日本語教育に一般的に用いられる表記と，コーパスの代表表記とは一致しない点が多い。

2.3.2.1. コーパスの漢字表記

形態論情報の漢字表記部分である語彙素や代表表記などを見ると，一般的には漢字表記でない語彙まで漢字に統一されている。これは，実際のテキストでは仮名や漢字などが混在しているが，コーパスを解析し，形態論情報を付与するにあたり，見出し語としての表記を統一させる必要があるためである。小椋（2006, pp.150-159）には，代表表記の付与基準について以下のように記されている。

[1] 代表表記には，書き起こしテキストの基本形[16]の表記を採用する。ただし，[2] 以下の規定に該当するものについては，その規定によって代表表記を定める。

［基本形］	［代表形[17]］	［代表表記］
コウモリ	コウモリ	コウモリ
表わす	アラワス	表わす

[2] 以下に挙げるものは，転記テキストの基本形の表記によらず，各規定に基づいて代表表記とする漢字表記を定める。

(1) 数字の代表表記は，すべて漢字とする。

［基本形］	［代表形］	［代表表記］
九十四	キュウジュウヨン	九十四

16 基本形とは，CSJ において漢字仮名を中心に可能性の高い形式で音声を文字化したもの。すなわち，解析する前のテキストの中の形を指す。
17 代表形とは，CSJ において単位認定基準に基づいて認定した各単位に与える見出し語情報で，片仮名で示されている。

(2) 助詞・助動詞の代表表記は，すべて平仮名とする。

［基本形］	［代表形］	［代表表記］
程	ホド	ほど

(3) 以下に挙げる語については，転記テキストにおける表記の使い分けにかかわらず，以下のとおり代表表記を統一する。

［基本形］	［代表表記］
現われる，表われる	現われる
替える，代える	替える
言葉，詞	言葉
箱，匣	箱
咄本，噺本	咄本
張る，貼る	張る
汎化，般化	汎化
平行，並行	平行
混ぜる，交ぜる	混ぜる
巡り会える，巡り合える	巡り会える
食料，食糧	食料
ゼロ，〇，零	ゼロ
戦う，闘う	戦う
取れる，捕れる	捕れる
引き延ばす，引き伸ばす	引き延ばす
柔らかい，軟らかい	柔らかい※

(4) 同一語で漢字表記と仮名表記との使い分けが行われている場合は，すべて漢字表記を代表表記とする。

［基本形］	［代表形］	［代表表記］
行く，いく	イク	行く

置く，おく　　オク　　置く

(5) 和語・漢語のうち，転記テキストの基本形において一部又は全部が平仮名で表記されているものについては，『岩波国語辞典』第 5 版（岩波書店）及び『国語大辞典』（小学館）を参照して，可能な限り漢字表記を当てる。

［基本形］	［代表形］	［代表表記］
あなた	アナタ	貴方
これ	コレ	此れ
無理やり	ムリヤリ	無理矢理
あるいは	アルイハ	或いは
おる	オル	居る
する	スル	為る
うまい	ウマイ	甘い

『岩波国語辞典』第 5 版及び『国語大辞典』を参照しても漢字表記を当てることができない，又は漢字表記を当てることが適当でない場合は，転記テキストの基本形の表記（平仮名表記）を代表表記とする。

［基本形］	［代表形］	［代表表記］
とんかち	トンカチ	とんかち
とんでも	トンデモ	とんでも
もう	モウ	もう
やや	ヤヤ	やや
にこやか	ニコヤカ	にこやか

[3] タグ(A) が付された記号等については，タグ(A) の右項に記載された表記を代表表記とする。

［基本形］	［代表形］	［代表表記］
(A エイチエムエム;ＨＭＭ)	エイチエムエム	ＨＭＭ

(A エスオメガエル;ＳωＬ)　　エスオメガエル　　ＳωＬ

(A ケー;Ｋ)　　ケー　　Ｋ

(A ニスト;ＮＩＳＴ)　　ニスト　　ＮＩＳＴ

(A エスアイドット;Ｓｉ．)　　エスアイドット　　Ｓｉ．

ラージ(A ラムダ;λ)　　ラージラムダ　　ラージλ

ラージ(A シー;Ｃ)　　ラージシー　　ラージＣ

アルファベット 1 文字が 1 短単位となる場合,タグ(A) の右項に小文字で表記されていても，代表表記は大文字にする。

［基本形］　　［代表形］　　［代表表記］

(A エス　　エス　　S

アイ;Ｓｉ)　　アイ　　I

以上が代表表記の基準である。このうち，日本語教育で一般的に用いられる漢字表記と異なり，特に問題になるのは[2]の(4)(5)である。例えば，代名詞を「此れ」「其れ」などと表記したり,「やはり」を「矢張り」,「うまい」を「甘い」などと表記したりすることは，よほど特殊な文脈でない限りありえない。また，国語辞典を参照して可能な限り漢字を当てるというような原則に則って表記を決めているため，常用漢字以外のものも多く見られる。

2.3.2.2. 日本語教育語彙表の漢字表記

それでは，日本語教育語彙表の漢字表記はどうなっているのであろうか。コーパスの形態論情報を語彙表作成に応用するにあたり，日本語教育語彙表における漢字表記とコーパスの代表表記との違いをまとめ，見出し語の表し方について考察する。

日本語教育語彙表の見出し語には，まず平仮名と片仮名などで記され，その語彙の意味を識別する手がかりとして漢字を付けるパターンのものがある。このタイプのも

のには，例えば，国立国語研究所（1984）の「国語研教育基本語彙」,「出題基準」,「品詞別 A~D レベル別 1 万語語彙分類集」などがある。この漢字表記に関する細かい規則は説明されていないため詳細は不明である。しかし，実際に語彙表を見ると次のようなことが分かる。まず,「なかなか」「かなり」「いらいら」のような副詞の類,「これ」「それ」などの指示代名詞,「しかし」「そして」などの接続詞などは，平仮名のみで漢字は併記されていないことが多いようである。また，食べ物，飲み物，動物，植物などのカテゴリーの語は，漢字表記があっても併記されないものも多い。このようなカテゴリーの語には漢字があっても常用漢字表外の漢字であったり，平仮名や片仮名で記すことのほうが一般的であったりすることも関係していると考えられる。

見出し語を平仮名や片仮名で示さず，漢字表記にしているリストもある。このタイプには，玉村（2003)「中級用語彙―基本 4000」がある。これも特に漢字表記の基準などは記されていないが，見出し語が平仮名や片仮名のリストと同様に，教育的配慮に基づく漢字表記になっている。

このように，日本語学習語彙表の漢字表記に関しては，特にどのような表記に統一すべきというはっきりとしたコンセンサスはない。特に，常用漢字表の漢字に制限するなどということも行われていないようである。常用漢字表外の漢字，常用漢字表および付表に示されていない音訓，当て字や熟字訓の類，常用漢字であっても平仮名表記が一般的なもの，などについても漢字表記が示されているものもある。このようなものについて括弧や符号を付けるなどの工夫がなされている「品詞別 A~D レベル別 1 万語語彙分類集」のような語彙表もある。

コーパスの漢字表記と，日本語教育語彙表の漢字表記との違いは，コーパスのほうが機械処理上の都合から「可能な限り漢字表記を当てる」というような基準を設けているのに対し，日本語教育語彙表では通常平仮名で使われる語については平仮名のままにするというような教育的配慮がされている点である。

しかし，日本語教育語彙表に使う漢字表記には明確な決まりはなく，語彙表によってもそれぞれであるのが現状である。これは，漢字の難易度と語彙の難易度が別のも

のとして考えられており，日本語教育語彙表は語彙の重要度や難易度を示すものだからである。コーパスの形態論情報を日本語教育語彙表に利用する場合はこの点に関して理解し，日本語教育的な表記を併記したり，そのまま使う場合は表記に関する情報を明記するなどの配慮が必要である。

2.4節　先行研究のまとめ

2.4.1. まとめ

第 2 章では，先行研究を概観した。ここでは，その概要を簡潔にまとめる。

日本語研究の分野では，従来，様々な分野の小規模な語彙調査が行われ，日本語教育基本語彙の選定には，そのデータが利用されてきた。また，語彙選定やレベル分けの方法は，専門家の方式が主流であった。そして，現在まで日本語教育語彙表の「定番」として最もよく利用されているのが「出題基準」である。しかし,「出題基準」は，最初に作成されてから 30 年以上が経過しており現在の語彙に対応しきれていないことや,「日本語能力試験」という特定の試験のために作られた語彙表であり，広く教育現場や研究目的で利用するには不十分であることなどが，問題点として指摘されていた。

一方，英語教育では，1964 年に Brown Corpus が公開されてから，基礎語彙の選定や辞書の作成などにもコーパスが活用されてきた。Brown Corpus 以降も様々なコーパスが整備され，日本語教育に比べて英語教育では大規模コーパスを利用した語彙表研究が進んでいる。

日本語のコーパスは近年整備が進められた。2009 年に BCCWJ の一部が公開され，これに準拠した日本語教育語彙表として松下（2011）や李・砂川（2012）が開発された。

その後， 2011 年に BCCWJ の完全版（総語数約 1 億語）が公開された。そして， この BCCWJ 完全版と話し言葉コーパスである CSJ に基づいて Tono et al.（2013）が作

られた。なお，これは頻度上位 5000 語までを掲載した頻度辞書で，日本語教育的観点からの重要度によってランキングしたり，レベル分けするなどして語彙を選定した基本語彙リストではない。

このように，日本語教育では古くから小規模の語彙調査などのデータに基づく基本語彙の選定が行われてきた。その中で，語の単位認定や表記の問題など，様々な日本語特有の問題についても研究が重ねられた。そして，2009 年以降は大規模コーパスが利用できるようになり，コーパス準拠の語彙表を作る動きが進んでいる。一方，コーパスの利用の仕方や作成方法は語彙表によって異なり，確立された方法論というものは存在しない。コーパスと統計指標に基づく語彙の選定に関わる研究は，日本語教育においては比較的新しいものである。

2.4.2. 問題の所在と本研究の意義

これまで，コーパス準拠の日本語教育語彙表として，松下（2011）や李・砂川（2012）が作成された。共に優れた研究成果であることは言うまでもないが，これだけあればどのような利用目的においても十分というわけではない。まず，この二つの先行研究では，使用したコーパス自体の均衡性や日本語教育語彙表を作成する元データとしての適切さなどについて，どのような検証を行っているのかが明らかではない。また，李・砂川（2012）は，語彙選定基準において統計的手法よりも専門家判定方式に重きを置いていることから，語彙選定の客観性を最重要視したものではないと考えられる。さらに，従来型の日本語教育語彙表の代表ともいえる「出題基準」を参照するなど，既存の日本語教育語彙表作成の方法論の延長線上にある部分もある。

しかし，コーパスに基づく基本語彙の選定においては，語彙表の目的に応じたコーパスバランスの検討，統計指標を活用した客観的語彙選定，教育的観点からの解釈と調整の 3 点が抑えられていることが重要である。このように語彙選定の客観性に主軸を置くことによって，従来型の語彙表とは質的に異なる語彙セットを取り出すことが

できると考えられる。

コーパスに準拠して語彙表を作成する場合，そのコーパスが語彙表の目的に合っているかを検討する必要がある（2.2.4.参照）。松下（2011）は，コーパスとして「書籍」と「Yahoo!知恵袋」からのデータを使用し，李・砂川（2012）は BCCWJ 2009 年度版領域内公開データの 10 媒体と独自に開発した日本語教科書コーパスを使用している。李・砂川（2012）は，日本語教育という目的に合わせて日本語教科書コーパスを加えており，ある程度コーパス自体への配慮が見られる。しかし，松下（2011）は，コーパス自体の均衡性や，サブコーパスのバランスが語彙表作成の目的に対して適切かどうかを考慮したという記述はない。

語彙の選定は，統計指標を基準として客観的に行えることがコーパス準拠の語彙表の利点である。分布統計のための指標は複数あるので（2.2.2.3.参照），目的に応じて統計指標を選択し利用するのが望ましい。松下（2011）は散布度として Juilland's D (Juilland et al. 1970)（2.2.2.3.参照）を用い，それに頻度を掛けた値を有用度として利用している。しかし，なぜ Juilland's D を使ったのかということについての説明はない。李・砂川（2012）は，分布統計は使用せず頻度を基準として語彙を選定している。

コーパスに基づき日本語教育基本語彙を選定する際には，語彙表の目的に合わせて日本語教育的観点からデータを解釈し，語彙のレベルや難易度を調整していく「味付け」のような作業が必要である。コーパスの出現頻度順に並べただけであれば，それは頻度表であり日本語教育語彙表としては不十分である。また，分布統計だけでランキングを確定するのも，語彙表の目的によっては十分とは言えない。しかしながら，従来型の語彙表のように，語彙の出現頻度や統計指標よりも専門家による判定を重視し主観的選定寄りになってしまうと，コーパス準拠であることの利点が失われてしまう。既存の語彙表を見てみると，レベル分けについては，松下（2011）は有用度指標を主な基準としつつも，「出題基準」を参考にレベルを確定している。しかし，「出題基準」は専門家判定方式によってレベル分けが行われたものである。李・砂川（2012）

は選定者の主観的判定を重視し，これによってレベル分けを行っている。

このように，既存のコーパス準拠の語彙表には，目的に応じたコーパスバランスの検討，統計指標を活用した客観的語彙選定，教育的観点からの解釈と調整の 3 点を満たしているものはない。しかし，これらはコーパスに基づき客観的に語彙を選定して語彙表を作成する上で，いずれも必要な手続きである。

語彙選定の客観性を重視し，既存の語彙表のものとは異なる語彙セットを取り出そうとするならば，まず，コーパス自体を十分に検討し，適切な統計指標を用いて語彙の重要度を数値化するべきである。そして，レベル分けは過去の語彙表の枠組みを利用したり，個人の主観に左右されやすい専門家判定方式に頼ったりするのではなく，統計指標を踏まえて分析した重要語彙を日本語教育的観点から再配列し，レベルを振り分けていくというのが，一つの有効な方法論として考えられる。

語彙表の作成方法は，その語彙表にどのような語彙が収録されているのか，すなわち，その語彙は何の指針になるのかという問題と直結する。例えば，教師がこのような語彙から学習すべきと考えるものなのか,学習者の語彙習得を反映したものなのか,または,「読む」「話す」など，ある言語活動において役に立つ語彙を客観的に選定したものなのか，などである。どのような語彙を選定するか，つまり，語彙表の目的によって語彙表作成の方法論は異なるであろうし，語彙表を利用する側もそれを理解している必要がある。

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―

第 3 章では，本研究で作成する日本語教育語彙表の作成方法について述べる。従来の日本語教育語彙表は，研究機関や教育機関が誰にでもアクセスできるわけではない語彙調査等のデータをもとに専門家の主観によって作成されるのが主流であった。先行研究ではコーパス準拠の語彙表も開発されはじめているが，語彙表の目的に応じてコーパスを評価したうえで利用し，統計指標によって語彙の重要度を定量化し，日本語教育的観点からの解釈で語彙の選定とレベル設定を行うという方法で作られたものはまだない（2.4.2.参照）。このようなことを背景に，本研究は，この三要件を満たす方法で日本語教育語彙表を作成することを目的とする。その作成方法の概要は図 1 の通りである。

まず 3.1.では，本研究で作成する日本語教育語彙表の目的と語彙表の設計について記す。本語彙表が対象とする学習者，総語数，表記や提示方法等についてはここで詳しく述べる。

3.2.では，教育語彙表作成に適したコーパスの選定と，最適化の方法について述べる。コーパスの最適化とは，コーパスのサブコーパスバランスが本研究で作成する語彙表の目的に合っているかを検討し，必要に応じてそのバランスを調整することを示す。コーパスの語彙頻度はそのコーパスを構成するテキストジャンルの特徴に影響を受ける。コーパスに基づく語彙表を作成にする際には，その語彙表の対象者や用途などの目的に合った語彙が抽出できるようにコーパス自体がデザインされていると，そこから得られる頻度情報が語彙選定において有益な判断材料となる。

3.3.では，3.2.で再構築したコーパスを頻度集計する。そして，複数の統計指標の中から本研究に適したものを選び，その計算方法について説明する。この統計指標によって語彙の重要度を定量化する。

3.4 では，語彙の難易度判定のための材料として単語親密度（天野・近藤，1999）

を用いる方法について述べる。単語親密度は日本語教育において語彙の難易度判定資料として使われることのある指標である。一方，日本語教育的な語彙の難易度と完全に一致するものではなく，単語親密度特有の傾向もある。そのため，この特徴が本研究の語彙表の目的と合わない場合には適宜調整も加える。

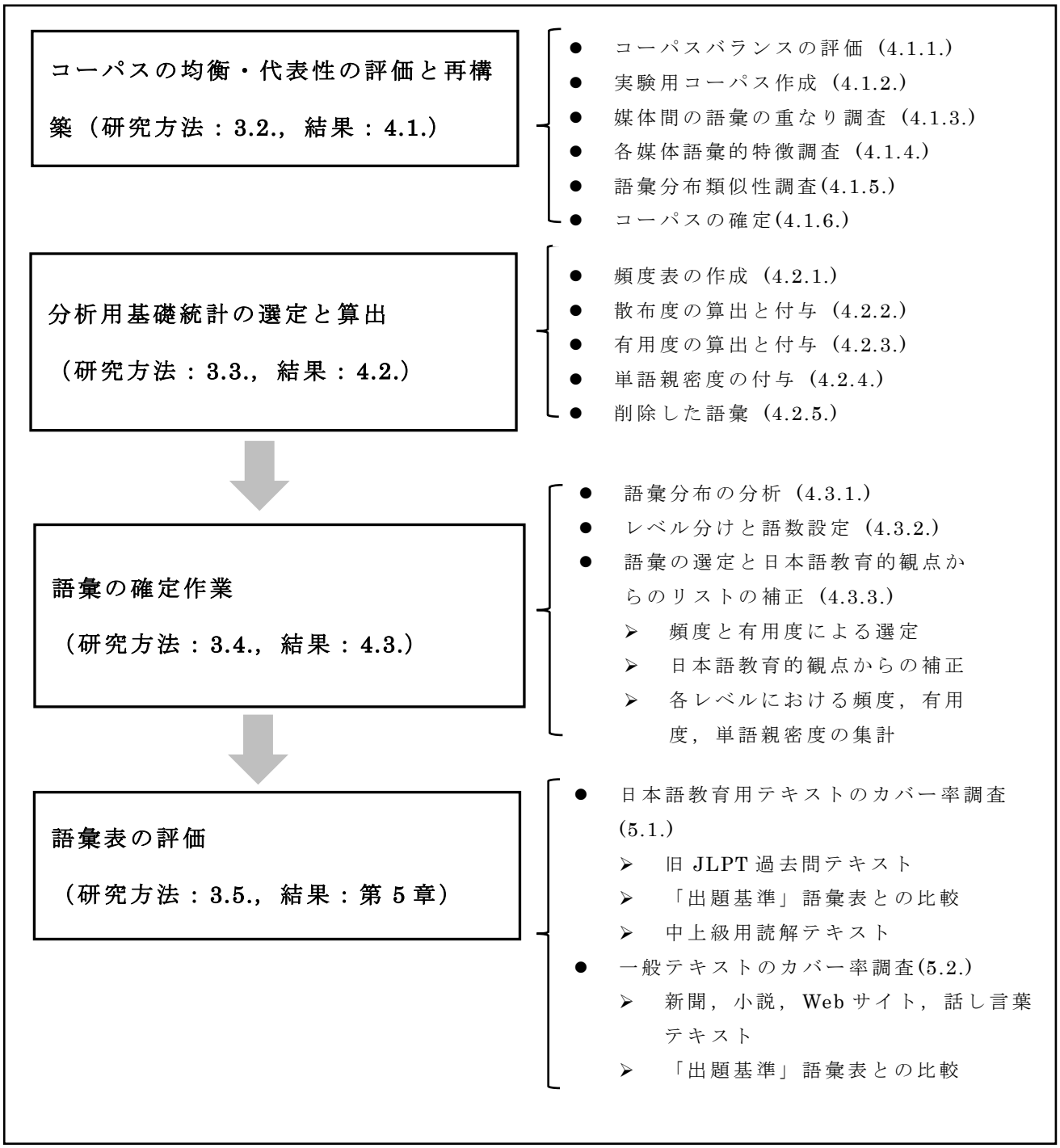

図 1　語彙表作成の方法

3.5.では，語彙の選定とレベル設定の方法について述べる。語彙のレベルの分け方すなわち，各レベルの語数設定は，データの特徴を頻度や分布の観点から分析して設定する。語彙の選定とレベル分けは，出現頻度，散布度，有用度により語彙の重要度を定量化し，単語親密度によって日本語教育的観点からの語彙の難易度を考慮したリストの再配列を行い，ランクを調整するという方法で行う。

3.6.では，語彙表の評価方法について述べる。ここではまず，日本語教育用に作られたテキストと，新聞，小説，Web サイトなどの日本語母語話者向けに書かれたテキストにおける語彙表の語彙のカバー率を調査する。日本語教育用に加工されたテキストと，一般向けのテキストの両方でテキストカバー率調査を行うことによって，初級から上級までの各レベルにおいて，語彙の選定とレベル分けが適切に行われているかを評価する。

さらに，日本語教育で最もよく利用されてきた従来型の語彙表の一つである「出題基準」と，本研究で作成する語彙表の語彙のテキストカバー率の比較を行う。本研究は可能な限り客観性を保った方法により，専門家判定方式による語彙表とは質的に異なる語彙セットを取り出すことを目指している。本研究の語彙表と，日本語教育における専門家判定方式による語彙表の代表格といえる「出題基準」とでは，テキストカバー率にどのような違いが出るかをここで検証する。

3.1節　目的（語彙表のデザイン）

3.1.1. 語彙表の総語数，対象者，利用範囲

本研究は，初級から上級の成人日本語学習者に向け，書き言葉の日本語を理解するために有用な語彙（受容語彙; receptive vocabulary）を選定した日本語教育語彙表を作成することを目的とする。以下，その総語数，対象者，利用範囲などについて具体的に説明する。本研究で作成する語彙表のデザインは表 14 の通りである。

総語数は 1 万語とする。この語数は,「出題基準」語彙表に倣ったものである。「出

題基準」語彙表の総語数は 1 万語で，初級から上級までを 4 レベルに分けて提示している（2.1.3 参照)。従来型の語彙表における語数は最大で 1 万語規模で，初級レベルから上級レベルまでをカバーするとされている。すなわち，従来から考えられてきた「上級学習者」に必要な語彙数は約 1 万語というコンセンサスがあったと見られる。一方，コーパス準拠で基本語彙を選定した李・砂川（2012）は約 1 万 8 千語を選定している。また，日本語能力試験改訂に伴い作成された語彙リストの語数は初級から上級，さらに「上級の上」の学習者を想定して約 1 万 7 千語が選ばれている。このような状況を見ると，従来型では 1 万語までを「上級レベル」とし，最近では 2 万語に近い語彙を「上級以上」の学習者向けに示す傾向のようである。

しかし，日本語教育語彙表に 2 万語に近い妥当な語彙を選定することは容易ではない。コーパス準拠で語彙表を作る場合コーパスの出現頻度が重要な選定材料となるが，中頻度以降の語彙に関してはコーパスが異なると頻度ランクも大きく変わってくる。Tono (2013)は複数のコーパスの語彙を出現頻度順にリスト化し，1 万語までを頻度順に 1000 語区切り（10 レベル； 1-1000, 1001-2000,...9001-10000）に分け，頻度順位の一致度を見たところ，最初の 1000 語ではコーパスによって頻度ランクの違いがあまり見られなかったものの頻度が低くなるほど違いが大きくなり, 5000 語以降あたりからはコーパスの違いが出現頻度の違いに大きく影響しはじめ, 9000 語から 1 万語の範囲では全部のコーパスに共通して出現する語彙自体が非常に少なくなるという現象が見られた。この結果は，コーパスの出現頻度を主な判断材料に 2 万語規模の語彙表を作ることの難しさを示している。

本研究で使用するコーパスは BCCWJ を日本語教育向けに調整するものである。しかし，コーパスのサンプリングの仕方が違えば，中頻度以降の語彙の重要度も異なってくることが予想される。一方，本研究の語彙表の目的は，初級から上級レベルまでの語彙を示すことである。そこで，本研究では，最近開発されている日本語教育語彙表と同様に 2 万語程度を示すのではなく，従来型の日本語教育語彙表のように 1 万語までを上級学習者に向けた最低限の語数として示す。

対象者は，初級から上級までの一般成人日本語学習者とする。これは，JSP（Japanese for Specific Purposes）や外国人児童生徒など，特定の目的のために日本語を学習している場合や特定の学習者層に向けたものとは異なるという意味での「一般成人日本語学習者」である。利用範囲は，主に語彙テストや語彙問題などの出題語彙リストとしての利用，教材作成における参考資料としての利用，語彙習得研究における利用などを想定している。本語彙表の見出し語配列は，頻度，散布度，単語親密度に基づいてレベル分けし，最終的な順位を付けた語彙をそのままランク順に配列する。また，使いやすさの面から，五十音順に配列したリストも付ける。本語彙表の項目には，見出し語，読み，レベル，品詞，語種，頻度（調整頻度），散布度，単語親密度，重要度ランクの情報を付与する。いずれも，本研究の利用範囲において有用な情報である。

表 14　本研究で作成する語彙表のデザイン

利用範囲	書き言葉の日本語を理解するための語彙を中心とする ①テスト利用 ②教材利用 ③研究における利用
対象者	一般成人日本語学習者
総語数	1 万語
見出し語の配列	ランク順（レベル順）・五十音順
付与する情報	見出し語 読み レベル 頻度（調整頻度：PMW） 散布度 単語親密度 語種 品詞 重要度ランク

3.1.2. 見出し語の単位と表記

本研究で作成する語彙表の見出し語の単位と表記は，コーパスを形態素解析した結果得られる「語彙素」に基づく。すなわち，語の単位は短単位（3 章参照）で示し，

表記に関しては漢字の難易度を考慮に入れないものとする。

語の単位に関する問題，つまり，どこまでを 1 語とみなすかという問題は，それだけで大きな研究テーマになる。日本語教育的に使いやすい単位というものは，教師の直観的には存在するが，それを体系化するという研究は進んでいないようである。また，本研究の目的は，見て理解できる語彙（受容語彙）を示すことである。そのような意味で，意味を持つ最小単位である形態素に近い単位まで区切ったものを見出し語の単位とすることは，本研究の目的に沿ったものであると考える。

また，表記に関しても，ここでは形態素解析の結果出力される語彙素の表記まま提示する。日本語教育においては，漢字の難易度についてもそれだけで大きなテーマであり，語彙の難易度とは切り離して考えられる。

本研究の目的は，語彙の難易度を示す日本語教育語彙表を作ることであり，漢字の難易度までを示すものではない。学習者が本語彙表をそのまま利用する場合は漢字表記にも配慮が必要になるが，教師が教材作成や研究目的で利用する際には語彙素表記のままで示すことに問題はなく，研究利用などにおいてはむしろ語彙素表記のままのほうが使いやすいとも考えられる。したがって，本研究における語彙表の見出し語の単位と表記は，形態素解析によって示される語彙素表記のままとする。

3.2節　コーパスの選定と最適化の方法の検討

3.2.では，本研究におけるコーパス選定の理由と，そのコーパスを再構築する必要性，および再構築の方法について述べる。

本研究では，日本語学習者が現代日本語書き言葉を読んで理解するために必要な語彙を選定することを目的とするため BCCWJ を利用する。BCCWJ を選んだ理由は，これが現在利用可能な現代日本語書き言葉均衡コーパスとして最大規模のものであり，サンプリング方法やコーパスバランスの面においても信頼性の高いコーパスだからである。

しかし，日本語教育語彙表作成のもととなるコーパスが母語話者の日本語を基準として作られたコーパスそのままでよいかという点には議論の余地がある。すなわち，既存のコーパスをそのまま利用して語彙表を作るのか，または，語彙表の対象者や目的に合わせてあらかじめサブコーパスバランスを調整したうえで，その語彙頻度や分布に基づき語彙を選定していくのか，ということを考えなければならない。それは，コーパスを構成するテキストジャンルやそのサブコーパスバランスが，語彙の頻度や分布状況を大きく左右するからである。

BCCWJ も複数のサブコーパスから成るが，テキストの分野・媒体によって語彙の出現頻度や分布は異なるため，その割合によって抽出される語彙が違ってくる。本研究では BCCWJ をそのまま使うのではなく，日本語教育語彙表作成に向けてテキスト媒体のバランスを最適化して利用する。

投野・本田（2016）では，BCCWJ からランダムサンプリングした書籍，国会会議録，白書，Yahoo!知恵袋の 4 分野（各 20 万語，合計 80 万語）について，語彙頻度と散布度を用いて比較を行い，どの分野のテキスト媒体にも安定して出現する語彙は，頻度上位 100 語程度までであることを明らかにした。このことからも BCCWJ を利用した場合，それを構成するテキスト媒体のバランスによって，語彙頻度表の頻度順位にはかなり違いが出ることがわかる。

BCCWJ は均衡コーパスであるが，必ずしも日本語学習者にとって必要なジャンルから抽出されたテキストで構成されているわけではない。例えば，白書や法律に関するテキストなどがある。このようなテキストが多く含まれるとその特徴語の頻度ランクが上位になり語彙選定やレベル分けに影響する可能性がある。

語彙表作成においては，コーパスの選定と再構築が重要となる。語彙表作成におけるコーパス利用について，投野・本田（2016）は以下のように述べている。

教育語彙表を構築する基礎となるコーパス・データの収集が語彙表作成の大きなポイントとなる。この場合，教育語彙表の使用目的に応じて，標本抽出（sampling），

代表性（representativeness），均衡（balance）という三つの概念を考慮することになる。大きな方法論として，コーパス構築の段階で綿密な設計を行い，分野バランスなども十分考慮に入れて構築されたコーパスの頻度をそのまま語彙表として採用するという方法と，コーパスからの頻度情報は一つの参考データとして用い，それ以外の様々な指標とからめて総合的に判断するため，コーパスの構築自体は比較的大まかな設計で行う，という二つのやり方がある。どちらの場合にせよ，「入れたものが出てくる」というコーパスの必然的な特徴を十分理解して，コーパスに含める資料の内容の吟味は適切に行われなければならない。

そこで，本研究では BCCWJ を利用するにあたり，前者の方法論（コーパス構築の段階で綿密な設計をする方法）をとり，BCCWJ のサブコーパスに含まれるテキスト媒体のバランスから検討していく。具体的には，BCCWJ から一部をサンプリングし，媒体間の語彙の重なりや，各媒体の語彙的特徴を観察し，クラスタ分析を用いて媒体の語彙分布の類似性を考察していく。そして，それらの結果を踏まえて BCCWJ のコーパスバランスを日本語教育語彙表作成という観点から再検討し，頻度や分布統計から日本語教育に合った語彙を抽出できるようなコーパスを必要に応じ特定のテキスト媒体の割合を減らすなどして再構築するという方法をとる。

3.3節　分析用基礎統計の調査と検討

ここでは，まず再構築したコーパス（BCCWJ）を形態素解析し，頻度集計し，不用語を削除する手続きについて述べる。そして，本研究で利用する統計指標等（散布度，有用度，単語親密度）を付与する方法について説明する。また，本研究の語彙表作成に当たり，これらの統計指標を選んだ根拠についても記す。

3.3.1. 頻度表の作成

ここではコーパスの頻度集計については，まず，3.2.でサブコーパスバランスを検討し再構築したコーパス（BCCWJ）を，形態素解析ツール「茶まめ」（Unidic-MeCab）を使って形態素解析する。次に，「茶まめ」によって形態素に品詞，語種などが付与されたデータを頻度集計する。このような方法でコーパスを頻度表の形にする。この段階で語彙表と直接関係のない記号や空白は除く。

3.3.2. 散布度の選定と計算方法

ここでは，本研究で利用する散布度を選定した根拠と，その計算方法について説明する。

散布度には複数の指標があるが，第 2 章でも述べたように，語彙の散布度を示す指標として，コーパス言語学では Juilland's D (Juilland et al., 1970) や Carroll's D_2 (Caroll, 1970) が古典的なものとして今でも利用されている（投野・本田，2016)。日本語教育語彙表の例では，松下（2011）が Juilland's D を，Tono et al. (2013)が Carroll's D_2 を使用している。しかし，いずれの例においても，なぜそのような統計指標を選択したのかについては明らかではない。一方，最近では，Gries (2008) が DP という指標を提案している。Gries (2008) は，DP を Juilland's D や Carroll's D_2 などの欠点を解消し，かつ，計算方法も容易で優れた指標としている。しかし，日本語教育語彙表の作成において DP が用いられた例はない。そこで本研究では，散布度を示す指標としてこの Gries's DP を利用する。

本研究における DP の算出方法は以下の通りである。まず，コーパス（BCCWJ）を 2 万語区切りのテキスト（サブコーパス）に分割する。そして，各媒体から 50 ファイルずつランダムサンプリングし（全体で各媒体 100 万語ずつをサンプリング)，媒体ごとの DP を求める。最後に媒体ごとに計算した DP の平均値を出し，全体の DP とする。

2 万語区切りのテキストに分割するのは, DP の計算をする際にサブコーパスのサイズが同じでなければならないという前提に基づく。そのため，すべてのテキストを同じサイズに分割する必要がある。また，分割するサイズは小さすぎても，大きすぎてもよくない。例えば，数百，数千という小さな単位でテキストを分割すると，1 万語を取り出すという目的においては必要以上に分布が安定しないことになり，ほとんどの項目について DP が高い値になる（分布の安定しない語としてみなされる）ことが予想される。本研究の目的は，教育語彙として 1 万語を選定することである。どのサブコーパスからも比較的安定して得られる 1 万語を抽出するための DP 利用なので, 1 万語に近い値で分割するのが適当である。また，本研究で作成する語彙表に選定する語彙は，内容語のみで機能語を削除するほか，数詞，指示代名詞，対概念の語彙なども削除する計画である。したがって，サブコーパステキストに含まれるこれらの語数についても考慮し，2 万語区切り程度のテキストにするのが適当であると考えた。

2 万語に分割したファイルを 50 ファイルずつ，媒体ごとに 100 万語をサンプリングしたのは，コーパス（BCCWJ）を構成するテキスト媒体ごとのサイズがそれぞれ異なり ，同じサイズに統一するには 100 万語が限度であったためである。本研究ではテキスト媒体ごとに散布度（DP）を計算し，その平均値をコーパス全体の散布度として利用する方針をとる。そのため，テキスト媒体ごとの語数はそろっていた方がよい。このような実用的な理由から, 各媒体 100 万語ずつをサンプリングし, 散布度（DP）の計算を行った。

3.3.3. 有用度指標

有用度指標（utility measure）は，頻度と分布統計を合成して算出するもので，従来の研究には Juilland’s usage coefficient U（Juilland, 1964）や Carroll’s Um（Carroll, 1970）などがある（第 2 章参照）。本研究では，散布度のほか，語彙の選定，レベル分けのために有用度も利用する。語彙の選定とレベル分けの際には統計指

標に閾値を設定して行うことになるが，頻度と散布度について別々の閾値を設定するのは難しい。なぜなら，高頻度で分布が安定する語彙もある中で，高頻度でも分布の安定しない語彙もあるからである。そのような場合，頻度と散布度それぞれの閾値を設定してレベル分けするのは困難で，作業も煩雑になる。

そこで本研究では，頻度と散布度を掛けてオリジナルの有用度指標とし，これを利用して語彙選定を行う。この基本的な考え方は，特徴語指標の一つである tf-idf に基づく。tf は ”Term Frequency”（単語の出現頻度），idf ”Inverse Document Frequency”（逆文書頻度）を示す。idf はその語の特定性を示す指標であり，頻度はその語の網羅性を示すものであるから，これらを単独で用いるよりも両方の性質を併せ持つように，二つの尺度を組み合わせてその語の重みを計算するという考え方が，tf-idf である（徳永，1999）。idf および tf-idf は以下によって定義される：

$$idf(t)=\log N/(df(t))+1$$

$$tfidf=tf \cdot idf$$

N は検索対象となる文書集合中の全体文書数，df(t)は索引語 t が出現する文書数を示す。idf はある索引語が少数の文書にしか出現しない場合に大きくなり，どの文書にも出現すると最少の値となる。

本研究では散布度の優れた指標として DP を利用するが，語彙選定においては恣意的な判断要素を最小限にするため，tf-idf の考え方に基づき，語の散らばりを示す散布度 DP に，語の網羅性を示す頻度を掛けて重みづけをしたものを，有用度指標として応用する。

ただし，DP は 0 以上 1 以下の値で示され，0 に近づくほど分布の安定度が高いことを示す指標であるため，DP をそのまま掛けるのではなく逆数にする。また，頻度は広範囲に及ぶので，値の変化を小さくするため対数にして DP の逆数と掛ける。このようにして有用度を計算し，データに付与する。

3.4節　教育的配慮としての単語親密度調査の利用

このようにして作成した基礎データに，語彙の選定の際の参考値として利用するため，Excel で『日本語の語彙特性』（天野・近藤，1999）の単語親密度を付与する。単語親密度は，日本語教育において語彙の難易度判定の一つの基準としてよく参照されている（徳弘，2005，川村，2006，押尾他，2008)。

単語親密度には，文字音声単語親密度，音声単語親密度，文字単語親密度の 3 種類の親密度がある。文字音声単語親密度は，文字と音声で同時に見たり聞いたりした場合における単語親密度，音声単語親密度は音声のみ，文字単語親密度は文字を見た場合のみの単語親密度である。本研究は，書き言葉の語彙の難易度を示す語彙表を作成することを目的としており漢字の難易度までを考慮したものではないので，原則として文字音声単語親密度を使用する。

ただし，文字音声単語親密度では，表記の難しさが影響して値が低くなる（なじみがないと判定される）ものが多く見られる。それらは特に常用漢字以外の漢字で示されているものや，一般的にはひらがな表記となるがここでは漢字（形態素解析結果の語彙素表記）で処理されている副詞などに多い。そして，そのような文字音声単語親密度は，音声単語親密度と比較すると大きな差があるのが特徴である。例えば，以下のような語彙が挙げられる：

○　常用外表記のため音声単語親密度より文字音声単語親密度が顕著に低い

「纏める（まとめる)」，文字音声単語親密度：3.469，音声単語親密度：6.062

「繋がる（つながる)」，文字音声単語親密度：4.219，音声単語親密度：5.844

○　一般的にひらがな表記のものがここでは漢字表記であるため文字音声単語親密度が音声単語親密度より顕著に低い

「唯（ただ)」，文字音声単語親密度：4.469，音声単語親密度：6.344

「可成（かなり)」，文字音声単語親密度：3.312，音声単語親密度：6.062

これらは見出し語表記の問題であり，語そのものの難しさ（意味の難しさ）とは直接関係のないものである。そして，単語親密度特有の傾向が語彙のレベル判定に直接影響することは，日本語教育的観点から語彙の難易度を検討するという本来の目的に反する。したがって，本研究では，このような項目に関しては文字音声単語親密度ではなく音声単語親密度を参照し，語彙の選定やレベル分けの際の参考値とする。

3.5節　語彙の選定とレベル分けの方法

3.5.1. レベル分けと語数設定の方法

語彙表の語彙のレベル分けと各レベルの語数設定は以下の方法で行う。本研究で作成する日本語教育語彙表は総語数 1 万語である。これは，コーパスによって中頻度以降の語彙の頻度ランクが著しく異なる（Tono, 2013）ことを受け，本研究のようにBCCWJ だけを材料にした場合，1 万語以上の語彙を選定することはあまり意味をなさないと判断したためである。したがって，本研究では上級レベルまでをカバーする最低限の語数として 1 万語を設定した（3.1.1.参照）。

ここでは語彙表を作成するにあたって，本研究で使用するコーパスの語彙分布の傾向を調査するため，頻度順位を 1 万語までを目安とし，1000 語区切りの単位で累積頻度や DP 値の散らばりを見る。さらに，1000 語区切りの各グループの頻度と DP の記述統計をもとにクラスタ分析し，グルーピングを行う。本語彙表のレベル分けの際の語数の決定はこのグルーピングの結果を参考に行う。そして，この出現頻度と語彙分布の傾向に基づいて頻度別に語彙をグルーピングした結果を分析し，本語彙表の語彙をどのようにレベル分けするか，また，各レベルを何語区切りにするかという問題について検討する。

3.5.2. 各レベルの語彙の選定方法

語彙の選定は，原則として有用度指標（3.3.3.参照）を基準に行う。具体的には，まず，3.3.3.の方法で算出した有用度を有用度の高い順に並べ替え，上位語から各レベルに設定した語数より多めの語彙を，そのレベルに入れる「候補語」として選定する。次に，その「候補語」の難易度を日本語教育的観点から調整するため，単語親密度を利用する。具体的には，有用度順に選定した「候補語」を単語親密度順に並べ替え，再び上位から各レベルに設定した語数分だけの語彙を残す。「候補語」に入りつつも，単語親密度による絞り込みでそのレベルから落ちた語彙は，一つ下のレベルの「候補語」として再び選定にかける。これが語彙選定方法の大枠である。

ただし，単語親密度には特有の傾向があり，本研究で作成する語彙表の語彙の難易度判定とは関わりのない要因，すなわち見出し語の表記等の問題で，単語親密度の値が極端に低く示される場合がある。このような場合は日本語教育的観点からの難易度を優先し，適宜調整を行う（3.4.参照）。

また，高頻度語彙についても，単語親密度の特徴によって重要語が選定から漏れないよう配慮する。例えば，「為る（する）」や「侭（まま）」は高頻度語で有用度の値も高いことが予想されるが，単語親密度の値は低い（なじみがないと判定される）。また，初級段階で文型として学習する「呉れる（くれる）」や「積り（つもり）」などについても同様である。これらの語彙は教育語彙表としては初級レベルに位置すべきものであるが，単語親密度によって調整を行うと中・上級レベルに移動することになる。したがって，コーパスの中核的部分を担う高頻度語彙に関しては，単語親密度よりも日本語教育的な難易度を優先し，語彙の選定とレベル設定を行う。詳しくは 4.3.3.2.で述べる。

3.6節　語彙表の評価方法

ここでは，語彙表の評価方法について説明する。本研究で作成する語彙表の評価と

して，日本語教育用に加工された読解テキストや，上級学習者が触れる可能性のある一般のテキストにおける語彙表の語彙のカバー率調査を行う。また，日本語教育で最もよく利用されている「出題基準」語彙表とのテキストカバー率比較も行い，考察する。

テキストカバー率調査は，日本語教育用テキストと，日本人向けに書かれた一般のテキストの 2 種類で行う。また，テキストカバー率調査は語彙表の各レベルについて行う。本研究で作成する語彙表は，助詞，助動詞，接辞等を抜いた内容語のみで構成されているため，ここで行うテキストカバー率調査は，これらを除いた内容語のみの重なりを見るものである。

日本語教育用テキストには，まず，旧日本語能力試験 1 級～4 級で過去に読解問題として出題されたテキスト本文を利用する。旧日本語能力試験はレベルが高い順に 1 級，2 級，3 級，4 級の 4 レベルに分かれている。1，2 級については 2005 年～2009 年の 5 年分，3，4 級については 2000 年～2009 年の 10 年分を使用する。3，4 級で 10 年分を利用するのは，1，2 級に比べてテキスト本文が短いためである。また，旧日本語能力試験読解問題以外には，国際交流基金が日本語教師向けに教材用素材や日本語教育情報を提供する Web サイト「みんなの教材サイト」に掲載されている「中級読解」「日本語教育通信エッセイ」を使用する。これについては 4.4.1.で詳しく述べる。

一般のテキストとしては，新聞，小説，Web サイトに加え，話し言葉コーパスも調査の対象とする。新聞は「毎日新聞 2010 データ集」（毎日新聞社）から，小説は「ポプラビーチ」（ポプラ社），「幻冬舎 Plus」（幻冬舎），「Web KADOKAWA」（角川書店）の三つの Web サイトより，実際に単行本として出版もされている作品から，Web サイトはブログ，日本語学習者向けのサイト，学校ホームページ，ショッピングサイト，Wikipedia から，それぞれ 1 万語ずつランダムサンプリングを行う。また，話し言葉コーパスは，「BTSJ による日本語話し言葉コーパス」（東京外国語大学・宇佐美真由美監修，2011）から「雑談」と「論文指導」の場面の会話をそれぞれ 1 万語ランダムサンプリングして使用する（5.2.1 参照）。

また，旧日本語能力試験読解問題テキストと，一般のテキストに関しては,「出題基準」語彙表と本研究の語彙表のカバー率比較を行う。テキストカバー率比較は各語彙表のレベルごとに行い，本研究で作成する語彙表の妥当性を検証する。

第4章　コーパスに基づく日本語教育語彙表の作成

4.1節　語彙表の基礎資料となるコーパスの選定と再構築

4.1.ではコーパスの再構築について述べる。本研究では，BCCWJ に基づき語彙表を作成するが，利用に際してはサブコーパスの分野バランスなども考慮に入れ，日本語教育語彙表作成向けにコーパスを再構築する（3.2.参照）。すなわち，BCCWJ には書籍，Web，白書，教科書，新聞，雑誌等の計 13 種類の媒体から構成されているが，その語彙分布にはそれぞれ特色があるのでそれらを分析し，サブコーパスのバランスを日本語教育語彙表作成の観点から調整する。そこで本章では，まず，BCCWJ のコーパスバランスの評価を行い，媒体間の語彙分布における相違性・類似性を観察し，各媒体の語彙的特徴について調査する。そして，これらの分析に基づきコーパスを再構築していく。

本章は以下の順に述べる。まず，4.1.1.では，BCCWJ の語彙分布について調査を行った投野・本田（2016）に基づき，媒体間の類似度や BCCWJ の語彙分布の安定性について考察し，コーパスバランスを語彙分布の観点から評価する。次に，4.1.2.では，さらに具体的なコーパス再構築の指針を得るため，BCCWJ から実験用コーパスとして一部をサンプリングした過程について述べる。4.1.3.では，4.1.2.でサンプリングした実験用コーパスを利用して，BCCWJ における媒体間の語彙の重なりについて調査する。4.1.4.では，4.1.3.の調査結果に基づき，他の媒体と語彙の重なりが小さい，すなわち，特徴語を多く含む媒体についてさらに詳しく見ていく。具体的には，各媒体特有の語彙とはどのようなものかを観察し，それらが教育語彙表作成という目的に合ったものかを検討する。4.1.5.では，クラスタ分析を用いて媒体間の語彙分布の類似性を統計的に分析する。最後に，4.1.6.では，分析結果に基づき，本研究の教育語彙表作成に向け，BCCWJ のコーパスバランスを最終的に調整していく。

4.1.1. BCCWJ のサブコーパスバランスの評価

投野・本田（2016）は，BCCWJ のコーパスバランスについて，Biber (1993) の方法に準拠し，国立国語研究所（2008)『BCCWJ 領域内公開データ』(2008・2009 年度版)の 10 媒体からそれぞれ 2000 語区切りのテキスト 100 ファイル(合計 20 万語)，計 1000 ファイル（合計 200 万語）をランダムサンプリングし，媒体ごとの語彙頻度の相関を分析した。その結果が以下の表 15～17 である。

表 15　BCCWJ 領域内公開データ 10 媒体 の頻度表（100 語）の順位相関

	LB	OB	OC	OM	OT	OW	OY	PB	PM	PN
書籍（LB）	1	0.88	0.86	0.76	0.79	0.58	0.84	0.97	0.88	0.73
ベストセラー（OB）	0.88	1	0.83	0.74	0.51	0.32	0.71	0.87	0.68	0.54
Yahoo!知恵袋（OC）	0.86	0.83	1	0.71	0.67	0.4	0.88	0.87	0.82	0.64
国会会議録（OM）	0.76	0.74	0.71	1	0.5	0.43	0.56	0.7	0.57	0.43
検定教科書（OT）	0.79	0.51	0.67	0.5	1	0.83	0.78	0.79	0.9	0.87
白書（OW）	0.58	0.32	0.4	0.43	0.83	1	0.56	0.57	0.72	0.83
Yahoo!ブログ（OY）	0.84	0.71	0.88	0.56	0.78	0.56	1	0.84	0.92	0.78
書籍（PB）	0.97	0.87	0.87	0.7	0.79	0.57	0.84	1	0.89	0.74
雑誌（PM）	0.88	0.68	0.82	0.57	0.9	0.72	0.92	0.89	1	0.89
新聞（PN）	0.73	0.54	0.64	0.43	0.87	0.83	0.78	0.74	0.89	1

表 16　BCCWJ 領域内公開データ 10 媒体の頻度表（500 語）の順位相関

	LB	OB	OC	OM	OT	OW	OY	PB	PM	PN
書籍（LB）	1	0.85	0.75	0.66	0.74	0.44	0.8	0.85	0.83	0.71
ベストセラー（OB）	0.85	1	0.81	0.58	0.59	0.25	0.86	0.86	0.82	0.62
Yahoo!知恵袋（OC）	0.75	0.81	1	0.62	0.63	0.37	0.88	0.83	0.83	0.66
国会会議録（OM）	0.66	0.58	0.62	1	0.61	0.62	0.62	0.68	0.68	0.65
検定教科書（OT）	0.74	0.59	0.63	0.61	1	0.63	0.64	0.72	0.74	0.71
白書（OW）	0.44	0.25	0.37	0.62	0.63	1	0.39	0.44	0.54	0.7
Yahoo!ブログ（OY）	0.8	0.86	0.88	0.62	0.64	0.39	1	0.85	0.86	0.72
書籍（PB）	0.85	0.86	0.83	0.68	0.72	0.44	0.85	1	0.88	0.71
雑誌（PM）	0.83	0.82	0.83	0.68	0.74	0.54	0.86	0.88	1	0.8
新聞（PN）	0.71	0.62	0.66	0.65	0.71	0.7	0.72	0.71	0.8	1

表 17　BCCWJ 領域内公開データ 10 媒体の頻度表（1000 語）の順位相関

	LB	OB	OC	OM	OT	OW	OY	PB	PM	PN
書籍（LB）	1	0.82	0.67	0.6	0.66	0.35	0.74	0.82	0.78	0.64
ベストセラー（OB）	0.82	1	0.73	0.53	0.54	0.17	0.8	0.81	0.76	0.55
Yahoo!知恵袋（OC）	0.67	0.73	1	0.52	0.56	0.28	0.82	0.75	0.76	0.57
国会会議録（OM）	0.6	0.53	0.52	1	0.52	0.61	0.53	0.62	0.63	0.61
検定教科書（OT）	0.66	0.54	0.56	0.52	1	0.52	0.56	0.67	0.67	0.63
白書（OW）	0.35	0.17	0.28	0.61	0.52	1	0.29	0.36	0.44	0.6
Yahoo!ブログ（OY）	0.74	0.8	0.82	0.53	0.56	0.29	1	0.79	0.81	0.67
書籍（PB）	0.82	0.81	0.75	0.62	0.67	0.36	0.79	1	0.84	0.64
雑誌（PM）	0.78	0.76	0.76	0.63	0.67	0.44	0.81	0.84	1	0.74
新聞（PN）	0.64	0.55	0.57	0.61	0.63	0.6	0.67	0.64	0.74	1

（投野・本田，2016，より）

上位 100 語までの高頻度語彙の場合には比較的どのテキストにも安定して出現する語彙が多いため全体に相関が高めであるが，500 語，1000 語とランクが下がるにつれて各分野のテキストの語彙特徴が表れてくる。これらは，フォーマルな書き言葉系（OW, OM）と話し言葉系（OY, OC）に大別されると見ることができる。

また，書籍・雑誌は全般に互いに相関が高い。これらの相関データからジャンルの傾向を分類すると，「書籍（LB, OB, PB／雑誌（PM)」,「ブログ（OY）／知恵袋（OC)」,「白書（OW）／国会会議録（OM)」の 3 グループに分けて考えることができる。したがって，この調査では，10 媒体の中にはある程度語彙分布が類似した媒体もあれば，異なる傾向を示す媒体もあることが示唆されている。

4.1.2. 実験用コーパスの作成

次に，この BCCWJ 語彙的傾向を媒体別にさらに詳しく見ていくため，BCCWJ の 12 分野[18]より新たに 20000 語区切りにしたデータを全媒体 50 ファイル（合計 100 万

[18] LB：書籍（流通実態），OB：ベストセラー書籍，OC：Yahoo 知恵袋，OM：国会会議録，OP：広報誌，OV：韻文，OT：検定教科書，OW：白書，OY：Yahoo!ブログ，PB：書籍（生産実態），PM：雑誌，PN：新聞の 12 分野。法律（OL）は頻度集計した際，極端に異なり語数が少なく，分析に含めると他の媒体のサンプリング語数も極端に減らさなければならないため，分析の信頼性が下がるという問題が生じる。また，高頻度語彙として挙がる語彙も特殊なものであり，明らかに他の媒体とは異質であった。法律に関連する特殊な語彙を基本語彙の中に含めることは本研究の目的とも合わないと考え，この媒体は初めから除外するこ

語）ずつランダムサンプリングし，実験用コーパスを作成した。ただし，総語数が 100 万語以下の媒体もあり[19]，実験用コーパス規模は全体で，ファイル数が 557，総語数 11,140,000 語（短単位）となった（表 18）。

なお，投野・本田（2016）では 2000 語区切りのデータで語彙頻度や分布について調査したのに対し，20000 語でサンプリングしているのは，本研究で作成する語彙表が 10000 語規模のものであるためである。実際に選定する語数に近い値で頻度や分布を見ていくことにより，本研究にとって有益な示唆を得られると考えた。また，本研究で作成する語彙表に最終的には含めない品詞（助詞，助動詞，接辞，固有名詞など）もこの中には含まれるので，10000 語よりも多い 20000 語でサンプリングを行った。

表 18　実験用コーパス

媒体名（略称）	語数 （短単位）	ファイル数 （20000 語区切り）
書籍（LB）	1,000,000	50
ベストセラー（OB）	1,000,000	50
Yahoo!知恵袋（OC）	1,000,000	50
国会会議録（OM）	1,000,000	50
広報誌（OP）	1,000,000	50
検定教科書（OT）	920,000	46
韻文（OV）	220,000	11
白書（OW）	1,000,000	50
Yahoo!ブログ（OY）	1,000,000	50
書籍（PB）	1,000,000	50
雑誌（PM）	1,000,000	50
新聞（PN）	1,000,000	50
合計	11,140,000	557

4.1.3. 媒体間の語彙の重なり

4.1.2.では実験用コーパスを作成した。4.1.3.では，これについて媒体ごとに頻度表を作成し，頻度上位 1000 語までの媒体間の語彙の重なりを調査した。その結果が表 19 である。

ととした。そのため，13 媒体ではなく，12 媒体で検討を行った。
[19] PN（新聞）は 90 万語，OV（韻文）は 22 万語をサンプリングした。

表 19　上位 1000 語までの重なり（語数）

重なる媒体	全体語数	書籍 LB	書籍 PB	雑誌 PM	新聞 PN	ベスト OB	知恵袋 OC	会議録 OM	広報誌 OP	教科書 OT	韻文 OV	白書 OW	ブログ OY
1 種類	1785	31	65	64	90	95	166	116	235	158	478	176	111
2 種類	557	56	84	66	89	81	133	104	107	81	67	143	103
3 種類	279	64	70	58	63	85	66	76	69	73	51	86	76
4 種類	214	78	73	77	74	81	62	78	59	70	40	85	79
5 種類	164	92	75	78	83	68	52	70	67	68	28	80	59
6 種類	138	104	76	87	77	70	57	77	61	61	29	73	56
7 種類	123	101	83	101	81	83	68	65	51	67	32	48	81
8 種類	83	77	75	73	58	62	46	57	44	51	23	44	54
9 種類	99	95	97	95	88	79	68	75	49	72	41	47	85
10 種類	88	88	88	87	83	82	71	73	54	85	43	44	82
11 種類	104	104	104	104	104	104	101	99	94	104	58	64	104
12 種類	110	110	110	110	110	110	110	110	110	110	110	110	110
総計	3744	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000

この結果を見ると，全ての媒体（12 種類）に出現する語彙は 110 語のみであった。「重なる媒体」の数が少ない部分の語数が多いほど，その媒体特有の語彙を多く含むことになる（重なる媒体数が 1 種類というのは，どの媒体とも重ならないということを意味する)。表 19 を見ると，書籍（LB，PB)，雑誌（PM）などは他の媒体と重なる語彙が比較的多く，韻文（OV)，広報誌（OP)，白書（OW)，Yahoo!知恵袋（OC)，教科書（OT)，Yahoo!ブログ（OY）は 1 種類のみに出現する語数が多い。

ただし，韻文（OV）に関しては，総語数が 22 万語しかないうちの上位 1000 語であるため，単純な比較はできない。そこで韻文（OV）を除き，さらに 2000 語から 10000 語までについて同様の方法で重なりを調べた（表 20～28)。

投野・本田（2016）が指摘しているように，高頻度語彙の場合には比較的どのテキストにも安定して出現する語彙が多いため，上位 1000 語ではその違いはあまり大きくないものの，1000 語区切りに上位 10000 語まで見ていくと，媒体の特徴が徐々に明らかになる。2000 語以降 10000 語までの間では，ベストセラー（OB)，Yahoo!知恵袋（OC)，国会会議録（OM)，広報誌（OP)，検定教科書（OT)，白書（OW）の 6 媒体が，他の媒体との重なりが少ないという結果が示された。

表 20　上位 2000 語

重なる媒体	書籍 LB	書籍 PB	雑誌 PM	新聞 PN	ベスト OB	知恵袋 OC	会議録 OM	広報誌 OP	教科書 OT	白書 OW	ブログ OY
1 種類	123	166	158	139	270	316	273	397	312	344	185
2 種類	138	147	118	155	158	218	221	178	182	291	176
3 種類	118	145	125	145	146	142	152	138	131	171	165
4 種類	160	130	150	161	159	149	160	145	117	155	170
5 種類	171	148	178	157	143	139	123	124	117	132	153
6 種類	178	137	159	150	142	112	114	114	123	109	126
7 種類	167	167	159	153	118	109	113	103	131	102	127
8 種類	182	197	187	180	144	131	139	122	155	115	160
9 種類	185	183	186	181	162	140	136	134	165	90	166
10 種類	235	237	237	236	215	201	226	202	224	148	229
11 種類	343	343	343	343	343	343	343	343	343	343	343
総計	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000

表 21　上位 3000 語

重なる媒体	書籍 LB	書籍 PB	雑誌 PM	新聞 PN	ベスト OB	知恵袋 OC	会議録 OM	広報誌 OP	教科書 OT	白書 OW	ブログ OY
1 種類	244	271	247	216	**454**	**471**	**471**	**492**	**483**	**502**	284
2 種類	187	211	180	230	239	308	325	299	256	426	263
3 種類	185	216	185	225	190	224	241	242	217	298	207
4 種類	192	177	199	203	177	178	189	189	155	199	230
5 種類	251	223	262	231	212	202	179	182	190	202	251
6 種類	242	206	243	228	212	200	186	162	161	153	215
7 種類	257	253	243	241	203	172	175	190	181	164	210
8 種類	289	289	278	273	225	197	190	183	243	154	231
9 種類	271	271	280	274	237	218	183	210	248	152	248
10 種類	335	336	336	332	304	283	314	304	319	203	314
11 種類	547	547	547	547	547	547	547	547	547	547	547
総計	3000	3000	3000	3000	3000	3000	3000	3000	3000	3000	3000

表 22　上位 4000 語

重なる媒体	書籍 LB	書籍 PB	雑誌 PM	新聞 PN	ベスト OB	知恵袋 OC	会議録 OM	広報誌 OP	教科書 OT	白書 OW	ブログ OY
1 種類	369	358	335	295	610	616	640	650	639	680	370
2 種類	262	299	269	300	335	405	425	359	337	517	358
3 種類	258	268	244	291	248	280	343	314	263	406	250
4 種類	236	290	287	282	240	296	246	270	242	300	295
5 種類	287	252	282	276	241	228	249	241	236	252	291
6 種類	313	278	325	296	269	257	232	216	224	201	305
7 種類	348	323	336	346	283	243	236	253	256	215	311
8 種類	315	319	314	298	261	235	197	221	253	174	277
9 種類	421	418	418	422	367	324	285	323	376	226	371
10 種類	444	448	443	447	399	369	400	406	427	282	425
11 種類	747	747	747	747	747	747	747	747	747	747	747
総計	4000	4000	4000	4000	4000	4000	4000	4000	4000	4000	4000

表 23　上位 5000 語

重なる媒体	書籍 LB	書籍 PB	雑誌 PM	新聞 PN	ベスト OB	知恵袋 OC	会議録 OM	広報誌 OP	教科書 OT	白書 OW	ブログ OY
1 種類	487	446	422	368	**787**	**751**	**811**	**830**	**806**	**819**	495
2 種類	323	385	331	382	404	516	528	439	425	643	416
3 種類	327	338	320	367	330	374	403	372	342	474	340
4 種類	320	323	345	351	282	331	350	343	290	381	344
5 種類	360	357	363	362	304	305	289	314	295	335	361
6 種類	345	332	370	354	299	294	276	266	274	269	341
7 種類	409	384	430	397	350	310	266	310	324	258	398
8 種類	428	434	419	418	360	309	296	291	333	251	381
9 種類	508	506	507	510	438	407	340	385	452	268	458
10 種類	544	546	544	542	497	454	492	501	510	353	517
11 種類	949	949	949	949	949	949	949	949	949	949	949
総計	5000	5000	5000	5000	5000	5000	5000	5000	5000	5000	5000

表 24　上位 6000 語

重なる媒体	書籍 LB	書籍 PB	雑誌 PM	新聞 PN	ベスト OB	知恵袋 OC	会議録 OM	広報誌 OP	教科書 OT	白書 OW	ブログ OY
1 種類	597	555	505	467	**942**	**927**	**1001**	**972**	**971**	**975**	610
2 種類	426	478	411	485	525	632	526	697	805	504	471
3 種類	409	415	420	458	413	452	440	455	544	414	434
4 種類	378	377	418	408	336	357	388	394	425	338	397
5 種類	396	405	419	406	338	367	360	326	374	343	426
6 種類	437	422	434	404	348	356	357	336	331	330	415
7 種類	461	441	485	467	408	349	349	334	335	380	443
8 種類	481	480	491	471	409	365	332	327	292	399	433
9 種類	583	590	589	601	519	468	468	401	322	522	553
10 種類	692	697	688	693	622	587	639	618	461	655	678
11 種類	1140	1140	1140	1140	1140	1140	1140	1140	1140	1140	1140
総計	6000	6000	6000	6000	6000	6000	6000	6000	6000	6000	6000

表 25　上位 7000 語

重なる媒体	書籍 LB	書籍 PB	雑誌 PM	新聞 PN	ベスト OB	知恵袋 OC	会議録 OM	広報誌 OP	教科書 OT	白書 OW	ブログ OY
1 種類	703	672	599	545	**1077**	**1083**	**1122**	**1173**	**1148**	**1151**	727
2 種類	529	531	525	546	656	738	779	613	580	861	578
3 種類	456	476	494	546	459	532	531	525	474	660	520
4 種類	497	496	485	541	430	437	476	481	436	509	484
5 種類	453	437	455	465	407	397	395	384	366	418	473
6 種類	475	497	527	478	396	417	392	414	402	396	472
7 種類	512	499	508	505	422	376	400	389	410	392	466
8 種類	571	573	573	543	496	437	385	403	472	337	522
9 種類	655	660	670	672	562	539	448	532	592	368	611
10 種類	766	776	781	776	712	661	689	703	737	525	764
11 種類	1383	1383	1383	1383	1383	1383	1383	1383	1383	1383	1383
総計	7000	7000	7000	7000	7000	7000	7000	7000	7000	7000	7000

表 26　上位 8000 語

重なる媒体	書籍 LB	書籍 PB	雑誌 PM	新聞 PN	ベスト OB	知恵袋 OC	会議録 OM	広報誌 OP	教科書 OT	白書 OW	ブログ OY
1 種類	864	778	684	613	**1239**	**1278**	**1324**	**1307**	**1307**	**1317**	844
2 種類	582	605	566	639	755	805	871	683	675	951	682
3 種類	523	557	590	642	548	599	623	608	564	745	589
4 種類	540	554	587	611	474	510	506	533	461	607	549
5 種類	562	535	528	575	487	480	474	469	473	501	551
6 種類	528	548	581	527	451	458	433	462	441	420	527
7 種類	561	581	575	564	485	444	449	460	467	437	535
8 種類	627	631	652	627	506	467	459	469	522	424	568
9 種類	760	755	779	740	672	615	513	620	676	434	708
10 種類	885	888	890	894	815	776	780	821	846	596	879
11 種類	1568	1568	1568	1568	1568	1568	1568	1568	1568	1568	1568
総計	8000	8000	8000	8000	8000	8000	8000	8000	8000	8000	8000

表 27　上位 9000 語

重なる媒体	書籍 LB	書籍 PB	雑誌 PM	新聞 PN	ベスト OB	知恵袋 OC	会議録 OM	広報誌 OP	教科書 OT	白書 OW	ブログ OY
1 種類	962	883	773	731	**1386**	**1427**	**1497**	**1439**	**1494**	**1538**	983
2 種類	669	674	660	711	830	935	959	794	768	1075	775
3 種類	644	648	672	709	657	681	743	678	659	814	676
4 種類	559	598	652	670	542	600	530	589	519	630	639
5 種類	645	623	606	648	533	510	543	555	504	571	577
6 種類	634	614	672	627	540	525	498	525	513	499	617
7 種類	626	654	635	613	548	513	478	495	516	460	594
8 種類	672	705	709	699	559	500	526	543	589	486	620
9 種類	851	852	874	841	740	694	611	708	739	531	803
10 種類	983	994	992	996	910	860	860	919	944	641	961
11 種類	1755	1755	1755	1755	1755	1755	1755	1755	1755	1755	1755
総計	9000	9000	9000	9000	9000	9000	9000	9000	9000	9000	9000

表 28　上位 10000 語

重なる媒体	書籍 LB	書籍 PB	雑誌 PM	新聞 PN	ベスト OB	知恵袋 OC	会議録 OM	広報誌 OP	教科書 OT	白書 OW	ブログ OY
1 種類	1085	991	888	841	**1553**	**1514**	**1688**	**1573**	**1693**	**1709**	1103
2 種類	771	768	726	784	910	1065	1088	910	862	1173	877
3 種類	692	693	751	759	743	766	802	743	698	869	743
4 種類	648	694	729	760	621	668	629	660	620	697	702
5 種類	693	680	690	712	577	617	578	625	566	631	651
6 種類	690	671	710	710	588	566	548	578	531	590	676
7 種類	698	722	720	692	614	568	532	559	573	508	667
8 種類	759	807	789	773	634	587	566	622	657	535	711
9 種類	898	893	920	887	789	714	648	740	776	582	829
10 種類	1122	1137	1133	1138	1027	991	977	1046	1080	762	1097
11 種類	1944	1944	1944	1944	1944	1944	1944	1944	1944	1944	1944
総計	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000

ベストセラー（OB），Yahoo!知恵袋（OC），国会会議録（OM），広報誌（OP），教科書（OT），白書（OW）の 6 媒体において，1 種類にしか出現しない語彙の数はランクが 1000 語下がるごとに約 150 語ずつ増加し，上位 10000 語では 1500 語以上が媒体特有の語彙であることがわかる。さらに，2 種類に出現する語彙も含めて見ていくと，表 29 および図 2 のようになる。

表 29　1 種類および 2 種類の媒体のみに現れる語の数

	書籍 LB	書籍 PB	雑誌 PM	新聞 PN	ベスト OB	知恵袋 OC	会議録 OM	広報誌 OP	教科書 OT	白書 OW	ブログ OY
1 種類	1085	991	888	841	1553	1514	1688	1573	1693	1709	1103
2 種類	771	768	726	784	910	1065	1088	910	862	1173	877
1・2 種類合計	1856	1759	1614	1625	2463	2579	2776	2483	2555	2882	1980

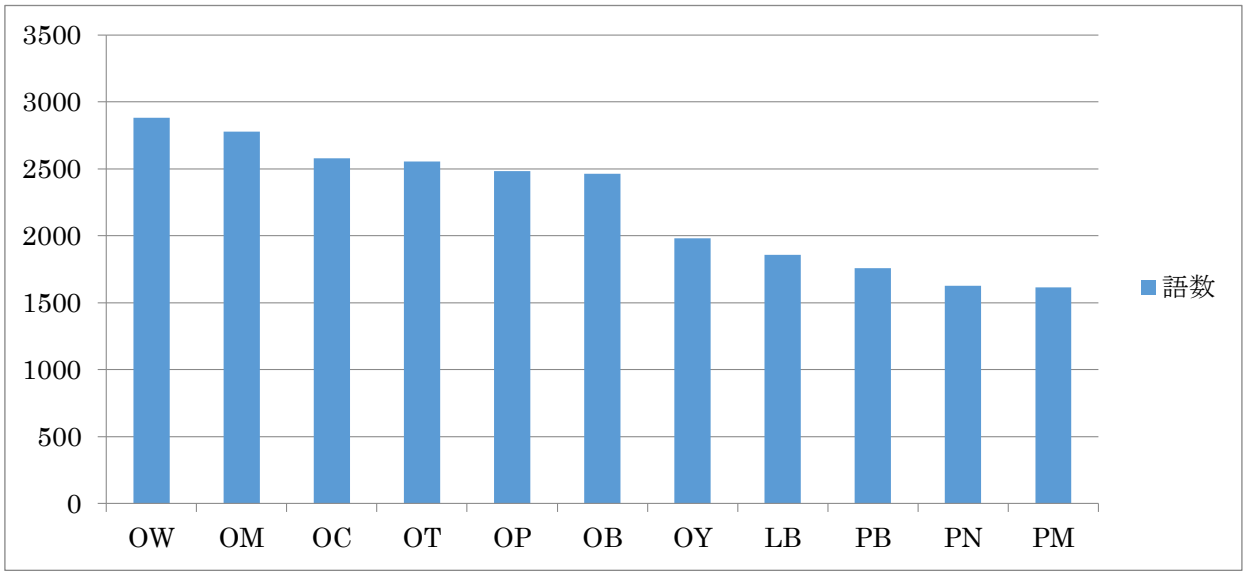

図 2　10000 語中 1 種類および 2 種類のみの媒体に現れる語彙の数

2 種類まで含めると，10000 語中，最大のもので 2882 語（白書，OW）までがその媒体に特徴的な語彙であると考えられる。全体の約 29%が特徴的な語彙である白書（OW）のほか，国会会議録（OM）（2776 語），Yahoo!知恵袋（OC）（2579 語），検

定教科書（OT)（2555 語)，広報誌（OP)（2483 語)，ベストセラー（OB)（2463 語）と続く。そして，少し間を置いて 2000 語以下では，Yahoo!ブログ（OY)（1980 語)，書籍（LB)（1856 語)，書籍（PB)（1759 語)，新聞（PN)（1625 語)，雑誌（PM)（1614 語）となっている。

このように，BCCWJ のうち特定目的サブコーパスに含まれる媒体は，語彙分布の特異性も他のサブコーパスと比べて顕著である。一方，書籍（LB と PB）および新聞（PN)，雑誌（PM）などは，それ以外の媒体と比べて他の媒体と共有する語彙が比較的多い傾向にあることが明らかになった。

4.1.4. 各媒体の語彙的特徴

4.1.3.の分析では，ベストセラー（OB)，Yahoo!知恵袋（OC)，国会会議録（OM)，広報誌（OP)，検定教科書（OT)，白書（OW）の 6 媒体の語彙が他の媒体との重なりが少ないという結果が示された。次に，この 6 媒体においてそれぞれ独自に現れる語彙（特徴語）はどのようなものか，上位 1000 語と 2000 語に絞って具体的に調べた。

表 30 から表 35 は，この 6 媒体それぞれの特徴語をまとめたものである。また特徴語であっても,日本語教育基本語彙として直観的に特に違和感がないものについては,参考として下線を付けた。以下，これらについて順番に示し，考察する。

(1) ベストセラー（OB）のみに出現する語彙

ベストセラー（OB）特有の語彙（表 30）には「喜ぶ，着く，弟，姉，父」等，の日本語教育基本語彙として直感的に違和感のない語彙が少なくない。小説中の登場人物名で用いられるためか，人名などの固有名詞も目立つ。なお，本研究で作成する語彙表には固有名詞は含めず作成過程で削除するため，語彙選定の結果には直接影響しない。

表 30　書籍：ベストセラー（OB）のみに出現する語彙

【上位 1000 語】 イエヤス イキ イサム イシカワ イノウエ エマ エリコ オダ カゼノ がる クウカイ ケンジ さあ じっと じゃ シンスケ すっかり タカハシ タクボク タミヤ で トクガワ とる ノブコ ノブナガ ヒデヨ ヒデヨシ ピン フユコ マサユキ ムロ ユダヤ ヨナイ ランコ リボン リュウホウ レイコ わし 挨拶 一瞬 噂 屋敷 仮令 回る 海軍 巻く 丸で 喜ぶ 宮 急 叫ぶ 近付く 刑事 見詰める 口調 皇子 姉 視線 出発 瞬間 証人 城 触れる 心 真 星 戦う 側 態度 大将 叩く 着く 弟 殿 逃げる 突然 婆 犯人 苗 夫人 父 父親 蕗 仏教 兵 癖 返す 返事 幕府 妙 立ち上がる 連中 呟く 頷く
【上位 2000 語】 ああ アキツ アコウ アサオ アサクラ アサノ アツヒメ アニータ アンジン アンドウ あんな イシダ イセ ウエスギ ウエノ おい オウミ オオイシ オーガズム オガタ オコウ オノデラ オヤマダ オリエ カサオカ カズノミヤ カツヨリ カマクラ カンスケ キシ きり キリノ キンダイチ グラス クロウ コウスケ コシジ コジマ サイチョウ サエコ サガ サザ サスケ サチコ サナダ シノブ スギコ スケキヨ セツコ ソウウン タカノ タクシー タケダ タツオ チャーリー テツヤ ど トキコ トヨトミ トラナガ どんな ナガシマ ニシゴウ はる ヒョウドウ ヒロコ フォーク ぶす ブラックソーン フルハシ へん ポケット マサシゲ マリコ マルクス ミチコ ミツナリ ミユキ ムラタ もの ユキムラ ヨウシチ ヨウハチ よし ヨシサダ ヨシダ ヨド ヨリトモ 愛情 偉い 一族 一旦 院長 右手 縁 下る 何しろ 何事 家臣 覚悟 巻き 館 顔色 喜び 奇妙 機嫌 気配 京 挟む 興奮 琴 稽古 警官 警部 功 慌てる 告げる 忽ち 此の頃 此奴 根本 些か 差し出す 砂 砂漠 使者 支店 死体 侍 守 首 衆 女史 女中 勝ち 匠 商事 商社 将軍 小屋 少佐 証拠 証言 冗談 嬢 新羅 親王 親切 親分 世 世間 正直 石 折 戦 戦闘 祖母 素直 相応しい 葬儀 息 袖 耐える 態々 大尉 大声 男の子 中将 注ぐ 弟子 提携 天下 途端 唐木 投げる 盗む 藤 肉体 入り口 任せる 念 燃える 馬 背後 泊まる 犯す 藩 微か 微笑 表 不意 不幸 浮かぶ 浮かべる 沸く 仏 陛下 傍ら 頬 本の 味方 密か 眠る 無い 命ずる 明かり 鳴る 黙る 尤も 唯 余程 様用 率いる 旅館 両 両手 力 礼 列車 廊下 匈奴

(2) Yahoo!知恵袋（OC）のみに出現する語彙

Yahoo!知恵袋（OC）の特徴語（表 31）には，「アレルギー，オイル，炎，外科，感染」などの日常語彙や具体物を表す語彙が多い。その中には「キッチン，トイレ，エアコン，洗濯，太る，石鹸」のような生活語彙，「カレー，キャベツ，牛乳，玉葱」のような食べ物名，「歯，薬，頭痛」のような体の部位や病気に関わる語彙など，基本的なものも多い。「食べる，飲む，買う，洗う，弱い」などの基礎語彙としても認定されうるようなものもある。したがって，Yahoo!知恵袋（OC）の特徴語には，日本語教育基本語彙として上位ランクに入るようなものが少なくない。

表 31　Yahoo!知恵袋（OC）のみに出現する語彙

【上位 1000 語】 アイス アドバイス アレルギー オイル カレー カロリー きちんと キャベツ クリーニング クリーム コーヒー こつ コンタクト コンビニ サラダ ジャガ芋 シャンプー スープ そんな ダイエット トイレ トマト パーマ パック ビタミン フライパン マッサージ マンション レシピ レンジ ローン 胃 医 引っ越し 炎 汚れ 温める 化粧 何方 可笑しい 我慢 外科 蓋 楽 乾く乾燥 干す 感染 慣れる 幾ら 吸う 牛乳 矯正 玉葱 近所 血液 見掛ける 元々 減る 庫 胡椒 交ぜる 喉 好み 砂糖 作る 雑誌 市販 試す 歯 捨てる 煮る 若しくは 弱い 主婦 臭い 柔らかい 汁 出し 出血 処方 少々 焼き 食べる 食べ物 水分 睡眠 整形 生える 生地 生理 生姜 石鹸 切れる 摂取 洗う 洗顔 洗剤 洗濯 掻く 痩せる 存知 多少 太る 体重 代わり 代謝 大根 脱毛 値段 知恵 恥ずかしい 中華 張る 直す 直る 賃貸 痛 痛み 痛む 通う 通販 店員 塗る 頭痛 独り暮らし 内科 鍋 入る 妊娠 濃い 納豆 買う 剥く 抜ける 半分 彼氏 皮膚 美容 付き合う 夫婦 婦人 布団 焚く 味噌 味付け 迷う 面倒 面皰 毛 薬 油 落とす 履く 離婚 溜まる 冷やす 冷蔵 冷凍 炒める 痒い 茹でる 鈍 鈍 黴
【上位 2000 語】 エアコン エステ オークション オーブン おでん おなら オリーブ カタログ カテ カテ違い カボチャ キッチン キムチ くっ付く クレンジング コレステロール コロッケ さっ さっぱり サプリメント ジーパン ジーンズ シチュー ジャム シャワー ジュース スーツ ストレッチ スプレー スポンジ タンク チキン ちん ティッシュ ドライ ドリンク ドレッシング ハード パウダー パスタ バナナ ハンバーグ ピーマン ビトン ピル ファンデ ファンデーション ブーツ プチ ブラシ フルーツ ペーパー ベランダ ポテト マスカラ マヨネーズ ミネラル ミルク ヨーグルト ライス ラップ レモン ワックス 慰謝 違法 違和 医院 芋 引っ越す 飲む 烏賊 炎症 汚い 汚れる 下地 下着 下痢 加減 家賃 果物 火傷 苛々 茄子 解凍 壊れる 外れる 咳 確か 顎 割れる 噛む 乾かす 寒天 勧め 完治 換気 缶 眼科 疑う 給料 牛蒡 禁煙 筋トレ 鶏 鶏肉 建て 現金 固まる 姑 後悔 御握り 御数 抗生 香水 合 今一 昆布 財布 擦る 子宮 指輪 脂 視力 歯磨き 似合う 耳鼻 煮込む 煮物 借金 手入れ 腫れる 寿司 収まる 縮毛 出来上がる 消化 消毒 焼酎 賞味 上手 蒸す 拭く 触る 食器 食材 食欲 尻 新築 浸ける 診察 身長 辛い 垂れ 炊飯 清潔 舌 洗浄 染み 染める 腺 素 素人 素麺 相続 増し 足す 多め 体型 台所 大抵 大分 宅 炭酸 炭水 鍛える 団子 知り合い 知人 中古 注射 虫歯 腸 通院 潰す 爪 剃る 天麩羅 唐辛子 豆乳 同居 独身 豚肉 内臓 軟らかい 匂う 日焼け 乳液 尿 濡れる 葱 悩み 排卵 剥がす 髪の毛 髪型 半身 半年 煩い 被る 微塵 貧血 不味い 敷く 浮気 腐る 風味 風呂上がり 服用 腹筋 沸騰 平気 片栗粉 便秘 崩れる 放し 放置 萌やし 飽きる 磨く 麻酔 毎回 鮪 味醂 眠る 名義 毛穴 目安 目薬 勿体 薬局 薬剤 有 余計 流行る 溜める 林檎 冷ます 冷める 脇 喘息 皺 瞼 蒲萄 蒟蒻

(3) 国会会議録（OM）のみに出現する語彙

国会会議録（OM）の特徴語（表 32）には，上位 1000 語の範囲で「伺う，申し上げる，仰る」（敬語動詞）のような日本語教科書的には初級後半レベルとされることの多い語彙が入っている。しかし，それ以外では，政治や経済に関する語彙が多く，直観的に初級レベルに合う語彙はほとんど見当たらない。

表 32　国会会議録（OM）のみに出現する語彙

【上位 1000 語】

ウイスキー カトウ クーリング サトウ しも スライド タガヤ どんどん なり ヒデオ フジワラ マルチ ヨシワラ ロッキード 異議 一千 引き上げ 営林 延長 恩給 下請け 画定 解消 海里 活 宜しい 救済 共 共済 境界 仰る 業界 局長 禁止 形 経過 計上 激甚 決議 結構 結論 見解 現行 戸 交渉 公社 公正 公団 公務 国鉄 国務 国有 根拠 伺う 斯様 施行 時点 次官 質疑 実情 若干 取り上げる 趣旨 十分 所管 商法 承る 承知 証券 慎重 申し上げる 正に 税制 先程 前提 相当 造林 存ずる 貸し付け 退職 台風 大蔵 大変 只今 棚 断層 中間 通産 提案 適当 電電 倒産 当局 当然 答申 答弁 入試 乃至 配分 伐採 不況 復旧 分科 方々 法案 本来 明確 融資 余り 要望 来 林野 論議 話し合い

【上位 2000 語】

アイザワ アイハラ アキヤマ アマヤ アルミ イサヤマ エトウ オオイデ オオキ オサム カジ カタヤマ カワサキ カワモト きちんと クボ サカタ さす サノ シオダ ショウゴ シロウ ジンイチ セールスマン セノ ソビエト タケウチ タケムラ ツクバ トクショウ トクダ トラブル ナカエ ナカガワ など ハヤシ ハラダ フサオ フルタ マージン マキ マサキ マジック マスオカ マツモト ミヤタ ムラサワ ムラヤマ メジャー モリタ ヤク ヤスハラ ヤタベ ヤバラ ヤマグチ ヨウスケ ヨシクニ 斡旋 遺憾 遺族 引き上げる 云々 鋭意 仮定 加算 介党 各省 割賦 勧誘 官庁 官房 敢えて 間 丸紅 期する 議題 詰める 逆鞘 協 強力 恐縮 偶々 傾斜 景 決壊 建て前 献金 絹 見地 遣る 現に 個々 公取 公平 鉱害 鉱業 鉱区 国債 国大 混信 砂防 採決 産炭 伺い 使命 司法 思い切る 支障 施業 私立 視察 試算 諮る 治山 治水 借家 主権 主査 取り締まる 手形 手元 手数 酒税 受け止める 衆議 従前 出す 出資 準ずる 順次 諸君 松 省令 常識 譲る 織物 食管 浸水 申し出 申し述べる 真剣 杉 政務 政令 正す 清酒 生糸 税率 積み立て 折衝 先般 選任 前向き 然り 然様 双方 層理 早急 増額 増税 造船 速やか 多額 妥当 代金 宅地 単価 担保 探鉱 探査 弾力 段々 値上げ 地質 町村 聴取 超過 直ちに 通達 紬 堤防 提起 提言 天災 展望 土砂 当面 等々 同意 特級 独禁 日韓 任意 能率 農政 売買 発注 発動 罰則 抜本 判定 反 彼処 被る 備蓄 品物 付帯 物品 並み 遍 俸給 法制 法的 冒頭 本年 本法 万全 民有 矛盾 有利 乱れ 率直 立法 立木 溜め池 了解 老齢

(4) 広報誌（OP）のみに出現する語彙

広報誌（OP）の特徴語（表 33）には，曜日名や「休日，祝日，毎月」のような初級レベルに合う語彙もある。しかし，「市政，税務，控除，庁舎」など，行政に関わる語彙なども多い。

表 33　広報誌（OP）のみに出現する語彙

【上位 1000 語】

アオイ イマバリ ウキョウ ウジ エコ カメヤマ キリシマ クサツ ゲンジ コウガ コウフ コーナー コミュニティー コンサート サカイ サポート シズオカ シミズ シャクジイ シン セタガヤ セミナー タカオカ タジミ ダンス テル テンノウジ トシマ ナゴ ネリマ ノト ハマキタ ハママツ ヒカリガオカ ヒダ ファックス プール フクイ プラザ ボランティア マチダ ヨドガワ リサイクル レジ ワジマ 案内 以内 印鑑 雨天 駅前 越す 往復 応募 下記 可 火 火曜 会員 会館 会費 改修 絵本 街 該当 活性 看護 願い 期日 気軽 記載 記入 休館 休日 救急 協働 教室 勤務 金 金曜 区民 係 啓発 掲載 軽減 健診 券 検診 見学 減額 後期 御覧 公民 口座 広場 広報 控除 講座 講師 講習 催し 祭 在学 在勤 在住 参画 子育て 市政 市税 市長 市内 支所 支部 氏名 資産 事前 持ち物 持参 舎 取り組み 手帳 受け付け 受け付ける 受講 収める 周年 就学 習慣 集合 祝 祝日 出張 順 初心 書類 商工 小 証 詳細 条例 触れ合い 振り仮名 振り替え 深める 申 申し込み 申し込む 申告 親子 診査 診療 人権 人数 水 水道 水曜 成人 正午 清掃 生う 生涯 税務 接種 先着 選考 全員 窓口 総務 送付 体育 体操 耐震 男女 知らせ 中止 抽選 駐車 丁目 庁舎 徴収 長寿 通知 定員 締め切り 土 当日 同伴 届け出 内線 日時 日程 乳幼児 入園 入場 認知 年始 年末 納税 納付 配布 発行 被 必着 票 表彰 不可 不要 扶養 保育 募集 母子 防犯 北部 本庁 毎月 毎週 毎年 未 民 明記 免除 木曜 問い合わせる 役所 有り 遊び 郵送 幼児 幼稚 葉書 要 立

【上位 2000 語】

木 丁 アスワ アリーナ イケブクロ イシカワ ウイング ウズマサ エプロン オオイズミ オリンピック ガイド カサハラ カスガ カミオカ カラスヤマ カワイ カワニシ キタザワ キチジョウジ キョウタンバ グラウンド クリーン コウナン コール コミセン コンクール コンテスト サークル サポーター サミット サロン サン シンドローム スルガ ダイセン タウン ダウンロード タマガワ タンバ チャレンジ テニス テンリュウ トーク トサ トマコマイ トヤマ ナカ パーク パトロール ファミリー フェア フェスタ フェスティバル フォーラム フルカワ フルカワチョウ ポスター ホンチョウ マーケット マナー ミズグチ ミズホ ムサシノ メタボリック モンゼン ヨーガ ライフ ルール ロビー ワーク ワークショップ ワチ 委嘱 育児 一時 一人一人 一斉 運行 営利 園児 遠慮 応急 下旬 下表 仮称 夏休み 花火 解散 絵画 開演 開館 開場 開通 街道 各回 学級 学年 楽 茅渟 観覧 還付 貴重 虐待 丘 救命 給食 京北 協同 教材 郷土 均等 区画 敬称 敬老 警報 決行 月額 月間 健やか 兼 献血 県民 県立 見直す 減量 古里 湖 公募 工作 広聴 行事 講 講義 高額 合唱 合同 国保 懇談 詐欺 催し物 在宅 参事 散策 市営 市道 試食 持ち 式 式典 実技 実践 写し 手話 種目 就労 修了 集い 集会 十字 縦覧 出店 出品 巡回 初級 所在 所定 書士 償却 小中学 消印 消火 上映 上旬 食育 心身 振り込む 震災 吹奏 水泳 随時 性別 生き甲斐 生き生き 生後 西部 青色 税額 赤 全額 爽やか 荘 贈呈 太鼓 滞納 台帳 大腸 卓球 託児 棚田 知的 中 中旬 昼食 町角 町内 訂正 添える 添付 転入 田 都合 土日 曜 陶芸 同館 特集 読み聞かせ 読書 届け 日頃 入会 入館 入賞 入門 任期 妊婦 納期 農園 配偶 発送 美 美化 筆記 標語 府民 分館 分室 分別 平日 別途 返信 便り 募金 本市 麻疹 万葉 満 民生 名簿 夜間 優秀 有料 柚子 預かる 擁護 来場 履歴 里 略 了承

(5) 検定教科書（OT）のみに出現する語彙

検定教科書（OT）の特徴語（表 34）には，「イオン，ナトリウム，プレート」のような専門用語が目立つものの，「英語，海，漢字，気候，分ける，通る」など，中級程度までのレベルには入っていてもよいと直観的に思われるタイプの語彙も含まれている。ただし，検定教科書（OT）は BCCWJ の中でも約 1%の割合しかない小規模の媒体なので，出現頻度においてはあまり大きな影響がないと思われる。

表 34　教科書（OT）のみに出現する語彙

【上位 1000 語】

イオン ギリシャ グラフ コイル ジュール ソフトウェア テープ ナトリウム プレート ベクトル ミリリットル メートル毎秒 衣服 英語 円 塩化 塩基 化合 化石 仮説 加速 火山 回転 海 各地 確かめる 角 楽器 割る 完成 漢字 管 観察 関数 岩 岩石 気候 気体 記号 鏡 曲線 極 形 結ぶ 結晶 元素 個体 語句 向き 工夫 恒星 考察 合成 国々 作曲 三角 酸 酸化 酸素 仕組み 四角 字 磁場 磁石 軸 質量 種子 周期 集 住まい 重り 重力 小数 乗 蒸気 食塩 振動 進 図形 垂直 水平 水溶 数値 制定 勢力 性質 成り立つ 整数 正 生かす 生ずる 盛ん 染色 蔵 測る 体積 堆積 対立 台車 大気 炭素 知識 地形 地層 地表 窒素 直線 通る 抵抗 伝わる 伝統 電圧 電流 当てる 等しい 統一 働き 動かす 動詞 導体 銅 特色 日常 濃度 波 配置 発芽 発条 発達 比例 肥料 付近 布 物体 分ける 分解 分布 分類 分裂 平行 平面 捕らえる 暮らす 方程 法則 縫う 棒 摩擦 密度 有機 溶ける 溶液 立体 粒子 力学 話し合う 惑星

【上位 2000 語】

アルミニウム アンペア エジプト オゾン カナザワ コジュウロウ コンデンサー シミュレーション ず スケッチ スペクトル セカンド っこ トルコ ナラ ニュートン ノア ノートハム パンフレット ビーカー プラスチック プレゼンテーション ベンゼン ホール マグマミシン メロス モル モル毎リットル モンゴル リズム リットル レポート ログ ワット 位 意 緯度 印 引き起こす 栄える 液体 鉛直 塩酸 塩素 王国 黄色 温室 下人 仮名 火星 解 解く 海溝 海水 殻 確率 楽章 掛け算 割り算 乾 還元 鑑賞 関わり 器官 基 気圧 記 貴族 教科 九々 屈折 傾き 形質 形容 桁 結び付く 結成 顕微 原始 古く 古典 固体 交わる 交雑 光 洪水 皇帝 酵素 高温 根 座標 作詞 山地 山脈 酸性 四辺 紫外 詞 字形 磁界 磁気 室内 実 斜面 主語 主題 修飾 順序 助 助詞 商人 小球 省 衝突 色紙 食物 振り子 浸透 身分 針 酢酸 水酸 数える 数学 正規 生き物 生存 西洋 石 石灰 積 赤道 節 素 組み合わせる 組成 相似 束 尊重 対数 対話 大正 大木 単 探究 探検 炭化 地殻 地中 地理 中性 中和 抽出 頂点 直径 直方 通過 定数 底面 適する 天体 纏め 電荷 電磁 電場 電離 都 唐 東西 当て嵌まる 等級 等式 動作 導線 道具 読み取る 日光 熱する 熱帯 波長 背 高 畑 発音 発酵 半ば 半径 半島 反射 板 飛び出す 筆算 紐 標本 復元 物差し 分かれる 分化 分数 平安 偏差 変形 編曲 放出 方形 泡立ち 望遠 本文 名詞 目盛り 遊牧 用具 落下 立法センチメートル 硫黄 硫酸 例題 列島 老婆 和歌 硼酸 胚葉

(6) 白書（OW）のみに出現する語彙

白書（OW）の特徴語は固い書き言葉を多く含み，上位 1000 語であっても日本語教育的観点からは直観的に上級レベルと考えられるような語彙が多い。また，白書は一般の日本語学習者が読む可能性がほとんどないような，特殊なジャンルのテキストである。コーパス全体に占める割合も低くはなく，白書（OW）の語数は一定量削減する必要があると考えられる。

表 35　白書（OW）のみに出現する語彙

【上位 1000 語】

ＯＥＣＤ アフリカ インフレ ウエート ウラン キロワット シェア ダム ニーズ プロジェクト やや レクリエーション 悪化 依然 依存 一層 衛星 円滑 援助 汚染 汚濁 卸売り 下回る 加盟 家計 果たす 箇年 回線 海域 海岸 海上 外交 各国 拡充 格差 確立 覚醒 活発 監視 危機 基金 基盤 寄与 機構 気象 急増 急速 近年 苦情 経常 継続 件数 健全 圏 検挙 原油 源 湖沼 港湾 高まる 高度 高等 合計 国費 債 採取 歳出 財源 指針 指数 支出 旨 資する 飼料 事犯 次いで 実質 実績 主 主要 取り締まり 需給 収益 収支 就業 集中 従業 従事 縮小 上回る 職業 食料 伸び 信頼 新規 進展 人員 推移 水産 水質 数量 性能 成果 西ドイツ 赤字 設立 節 先進 占める 船 船舶 相互 総会 総額 装備 騒音 増殖 増進 増大 対 大型 達する 達成 地盤 着実 注 著しい 沈下 賃金 通商 停滞 締結 適切 転ずる 電波 電力 途上 当該 当初 動向 動力 同期 内訳 日米 認可 燃料 濃縮 能 廃棄 背景 売り上げ 反映 犯罪 比率 非行 備考 品質 品目 普及 部会 変動 保安 保有 暴力 防火 本件 無線 名目 網 木材 輸送 抑制 立地 留まる 炉 齎す

【上位 2000 語】

ＡＳＥＡＮ カナガワ カンボジア ギガヘルツ キンキ サトシ サプライ シコク シンガポール セト チュウナンベイ パーライト パルプ ヒアリング ピーク プラント プルトニウム ブロック ベトナム マネー ミナマタ メガヘルツ モニタリング レート 移行 一貫 飲食 窺う 延べ 押収 横這い 欧米 下落 会期 海運 外食 各年 拡散 閣議 革新 学術 活力 慣行 換算 管制 簡易 緩やか 間接 丸太 基幹 基調 既存 起訴 休養 急激 救助 求人 牛 漁場 魚介 供与 協調 橋 局面 緊密 刑法 契機 軽水 堅調 検出 権原 顕著 元年 原 原型 原動 現況 戸数 故障 公営 公債 公衆 公立 広範 港 航行 高まり 高騰 合板 国営 国産 国定 採掘 採石 際する 在庫 罪名 作物 山村 産品 算出 暫定 仕入れ 四半 市況 志向 指向 指紋 死者 死傷 諮問 資機材 資材 自家 執行 疾病 質的 実 借り入れ 借款 若年 取り扱う 取り巻く 取り決め 取り纏める 種々 受け入れ 受刑 受注 受理 週休 集積 出所 純 順調 小型 省力 上陸 乗用 浄化 譲渡 食用 伸び悩む 伸び率 新型 新設 進学 進歩 人件 迅速 推計 水源 制約 精度 精密 製材 製紙 西側 西独 隻 先端 線量 船員 選定 前記 租税 総じて 総数 総量 遭難 増強 増減 造成 多角 多発 体様 対外 態勢 代替 台数 短 短期 短縮 団地 地価 畜産 蓄積 着工 着手 中核 著作 貯蔵 貯蓄 調和 鳥獣 直轄 通報 低減 低調 定住 定着 的確 鉄鋼 典型 伝送 電源 登山 度合い 投入 棟 討議 動機 同国 同年 鈍化 内海 内外 内政 肉用 年次 年報 農産 農用 農林漁業 配備 賠償 爆発 発効 半期 半数 犯 被曝 費目 必需 百貨 不可欠 不振 不法 付表 府 普賢 負債 負傷 幅広い 複合 分担 分配 並行 閉鎖 返品 歩行 歩道 補導 方策 防除 未然 無償 木造 目途 薬物 輸出入 優良 有害 有線 猶予 融合 予知 乱用 流出 流動 流入 旅客 両者 良好 緑地 林産 臨界 累計 累積 冷却 炉心 湾

4.1.5. 語彙分布の類似性

4.1.4.では媒体特有の語彙について上位 1000 語と 2000 語を例に見てきた。その結果，特有の語彙を多く含む媒体の中にも内容はそれぞれに傾向の違いがあり，ベストセラー（OB），Yahoo!知恵袋（OC），検定教科書（OT）のように教育語彙表作成において必要と思われるものと，白書（OW），国会会議録（OM），広報誌（OP）のようにバランスを調整すべきものとがあった。

そこで次に，このような傾向の違いをより客観的に分析するため，各媒体の頻度上位 1 万語をクラスタ分析し，語彙分布に関して媒体のグルーピングを試みた。

その結果，国会会議録やウェブの言葉を含む「話し言葉」的媒体のグループ（OM，OC，OY）と，書籍，新聞，雑誌，白書から構成される「書き言葉」グループ（LB，PB，PN，PM，OB，OW）に大別できた（図 3）。

さらに，その中の傾向の違いを見るため，表 36 のように全体を六つのグループに分けた。このように分けると，国会会議録（OM），広報誌（OP），白書（OW）は他の媒体との類似性が薄く，独立した傾向を持っていることがわかる。国会会議録（OM），広報誌（OP），白書（OW）は 4.1.3.でも明らかにしたように，上位 1000 語，2000 語の中でも独自の語彙を多く含む媒体である。また，その語彙の内容を見ても，教育基本語彙というよりも，専門性の高い語彙が高頻度語彙として挙がるのが特徴である。

また，ベストセラー（OB），Yahoo!知恵袋（OC），検定教科書（OT）も独自の語彙を多く含む媒体であるが，含まれている語彙を見ていくと，教育語彙として必要であるものも少なくないことが 4.1.4.で明らかにされた。クラスタ分析の結果でも，これらは特殊な専門語彙ばかりを含むタイプの媒体ではないことが示唆されている。ベストセラー（OB）は書籍（LB）と同グループに属する。Yahoo!知恵袋（OC）は Yahoo!ブログ（OY）と Web の語彙としての性質を共有している。検定教科書（OT）は新聞や雑誌などとも似た傾向があるようである。

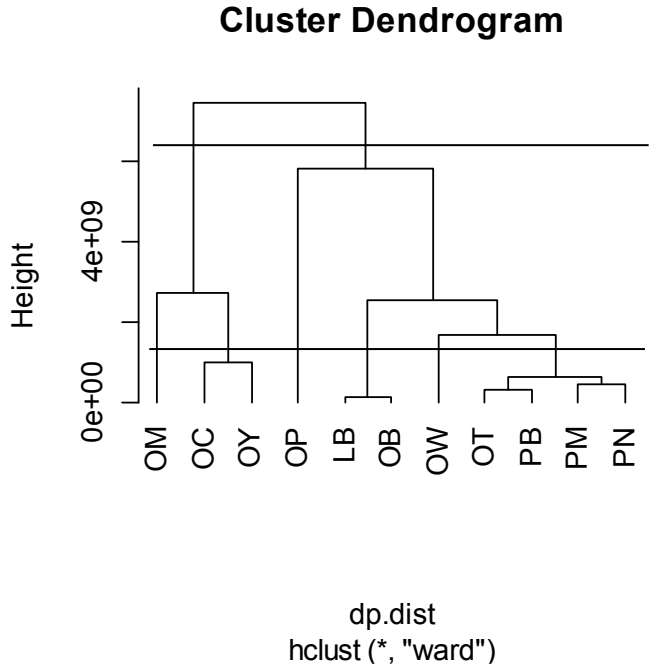

図 3　媒体のクラスタリング

表 36　媒体のグルーピング

話し言葉的	グループ１	会議録（OM）	固い話し言葉，政治経済の語彙
	グループ２	知恵袋（OC），ブログ（OY）	Web の言葉
書き言葉	グループ３	広報誌（OP）	自治体広報誌の語彙
	グループ４	書籍（LB），ベスト（OB）	書籍
	グループ５	白書（OW）	白書の語彙，固い書き言葉
	グループ６	教科書（OT），書籍（PB），雑誌（PM），新聞（PN）	教科書，書籍，新聞，雑誌

したがって，教育基本語彙として１万語を選定する際には，特に国会会議録（OM），白書（OW），広報誌（OP）の 3 媒体の特徴語が頻度順位などに大きな影響を与えないように調整する必要がある。一方，ベストセラー（OB），Yahoo!知恵袋（OC），検定教科書（OT）にも特有の語彙分布傾向があるが，国会会議録（OM），白書（OW），広報誌（OP）とは違った重みづけでコーパスバランスを調整したほうがよい。

4.1.6. コーパスの確定

本節では，ここまでの調査・分析結果を踏まえ，実際にコーパスバランスを検討し，確定する。基本的には全体から日本語教育的に重要ではない専門語彙の頻度が高い媒体の規模を縮小するという方法でバランス調整を行う。これまでの分析結果から，国会会議録（OM），白書（OW），広報誌（OP），およびベストセラー（OB），Yahoo!知恵袋（OC），検定教科書（OT）の 6 媒体が，このような媒体であることが分かった。しかし，国会会議録（OM），白書（OW），広報誌（OP）とベストセラー（OB），Yahoo!知恵袋（OC），検定教科書（OT）では，さらに違いがあることが分かった。

そこで，コーパスのバランスの調整では，ベストセラー（OB），Yahoo!知恵袋（OC），検定教科書（OT）のグループと，国会会議録（OM），白書（OW），広報誌（OP）のグループとで区別し，語数を削減する割合を検討した。また，高頻度語彙（1~2000 語）において不用な専門語彙の数が多いほど，その媒体の規模を縮小する割合も大きくなるようにする。BCCWJ の短単位データから，記号，空白等を除いた後のおよその総語数とコーパス中に占める割合は以下の通りである（表 37）。

表 37　BCCWJ の語数（短単位）の割合　※記号・空白削除後

媒体	語数（万語）	割合
書籍（PB）	2,855	27.2%
雑誌（PM）	444	4.2%
新聞（PN）	137	1.3%
書籍（LB）	3,038	29.0%
白書（OW）	488	4.7%
教科書（OT）	93	0.9%
広報紙（OP）	376	3.6%
ベストセラー（OB）	374	3.6%
Yahoo!知恵袋（OC）	1,026	9.8%
Yahoo!ブログ（OY）	1,019	9.7%
韻文（OV）	25	0.2%
法律（OL）	108	1.0%
国会会議録（OM）	510	4.9%
合計	10,493	100.0%

（国立国語研究所，2011, p.15 をもとに編集した）

頻度上位 1 万位までの中で，他の媒体との重なりがない語彙，および他の 1 種類としか重ならない語彙をその媒体「特有の語彙」とする。「特有の語彙」の割合は表 38 の通りである。

表 38　上位 1 万語中特有の語彙が占める割合

	特有の語彙の割合	平均との差
書籍（LB）	18.6%	3.8%
書籍（PB）	17.6%	4.7%
雑誌（PM）	16.1%	6.2%
新聞（PN）	16.3%	6.1%
ベストセラー（OB）	24.6%	-2.3%
国会会議録（OM）	25.8%	-3.5%
Yahoo!知恵袋（OC）	27.8%	-5.4%
広報誌（OP）	24.8%	-2.5%
検定教科書（OT）	25.6%	-3.2%
白書（OW）	28.8%	-6.5%
ブログ（OY）	19.8%	2.5%
平均	22.3%	0.0%

上位 1 万語までの中に特有の語彙を含む割合は，平均で 22-.3%である。それぞれの媒体の「特有の語彙の割合」が，どれだけ平均値（22.3%）から離れているかを示すのが「平均との差」である。この中で，その値が平均を上回っているのは，やはりベストセラー（OB）（2.3%），Yahoo!知恵袋（OC）（3.4%），検定教科書（OT）（3.2%），国会会議録（OM）（5.4%），白書（OW）（6.5%），広報誌（OP）（2.5%）の 6 媒体である。

媒体の規模を縮小する割合は，この「平均との差」を基準として定める。例えば，白書（OW）が含む特有の語彙は平均より 6.5%，広報誌（OP）は 2.5%多い。しかし，それぞれの総語数から 6.5%および 2.5%だけ削減しても，総語数に対して少なすぎるため頻度集計結果にはほとんど影響を与えない。そこで，恣意的な操作ではあるが，それぞれを 10 倍して白書（OW）は 65%，広報誌（OP）は 25%の割合で削減した。こうすることで特殊な媒体の語数を削減したことが全体の集計結果にも反映され，また，特殊な語彙が多い媒体ほど語数削減の割合を大きく，少ないものは小さくするという原則は守られる。

しかし，ベストセラー（OB），Yahoo!知恵袋（OC），検定教科書（OT）は特有の語彙が多数ありつつも他の媒体との類似性も認められる。また，特有の語彙が教育基本語彙作成という目的に適合しないものばかりではないので，国会会議録（OM），広報誌（OP），白書（OW）と同じように大幅な語数削減をせず，削減する割合を平均の差に対して 5 倍の値とした。すなわち，ベストセラー（OB）は 11%，Yahoo!知恵袋（OC）は 17%，検定教科書（OT）は 16%をその媒体全体の語数に対して削減した。これも，10 倍にしたときと同様に恣意的ではあるが，日本語教育的に不用な語彙を多く含む媒体は削減割合を大きく，特有な語彙が多くても日本語教育的に必要と感じられる語彙が多い媒体は少なめに削減するという原則は,少なくとも保つことができる。各媒体の語数を削減した結果が表 39 である。

表 39　コーパスから削減する割合と語数

	ベストセラー OB	国会会議録 OM	知恵袋 OC	広報誌 OP	教科書 OT	白書 OW
削る割合（%）	11%	54%	17%	25%	16%	65%
全体の語数（万語）	374	510	1,026	376	93	488
削る語数（万語）	43	277	177	94	15	316
残す語数（万語）	331	233	849	282	78	172

このような調整を加えると，コーパスのバランスは表 40 のようになる。コーパスの総語数は約 9463 万語になった。

表 40　調整前と調整後のコーパスバランス

	調整前		調整後	
媒体	語数（万語）	割合（%）	総語数（万語）	割合（%）
書籍（PB）	2,855	27.2%	2,855	30.17%
雑誌（PM）	444	4.2%	444	4.69%
新聞（PN）	137	1.3%	137	1.45%
書籍（LB）	3,038	29.0%	3,038	32.10%
白書（OW）	488	4.7%	172	1.82%
検定教科書（OT）	93	0.9%	78	0.82%
広報紙（OP）	376	3.6%	282	2.98%
ベストセラー（OB）	374	3.6%	331	3.50%
Yahoo!知恵袋（OC）	1,026	9.8%	849	8.97%
Yahoo!ブログ（OY）	1,019	9.7%	1,019	10.77%
韻文（OV）	25	0.2%	25	0.26%
法律（OL）	108	1.0%	0	0.00%
国会会議録（OM）	510	4.9%	233	2.46%
合計	10,493	100.0%	9,463	100.00%

4.2節　分析用基礎統計の算出

4.2.では，4.1.で再構築したコーパスに基づき頻度表を作成し，散布度，有用度，単語親密度などの情報を付与し，語彙リストには含めない語彙を削除する。

本章は以下の順に論じる。4.2.1.では，頻度表の作成方法について記す。4.2.2.では散布度，4.2.3.では有用度の計算方法とその値を付与する方法について説明する。4.2.4.では，単語親密度の付与について述べる。単語親密度における語の単位や表記は，形態素解析結果の語の単位（短単位）や表記（語彙素の表記）と一致しない部分もあるので，それらについて行う補正についてもここで説明する。4.2.5.では，本研究で作成する語彙表からは削除する語について述べる。

4.2.1. 頻度表の作成

ここでは頻度表を作成する方法について述べる。4.1.で BCCWJ を再構築したコーパスから未知語，空白，記号等を除いた後の総語数は，延べ語数で約 9463 万語，異

なり語数で約 19 万語であった（表 41）。このコーパスを，形態素解析ツール「茶まめ」（Unidic-MeCab）で形態素解析し，Unix コマンドやテキストエディタなどを用いて頻度集計して頻度表を作成した。

次に，この頻度表を頻度で降順に並べ替え上位 5 万語に絞った。これは効率的に作業を行うためである。本研究で選定するのは 1 万語の基本語彙であり，少なくとも頻度ランク 5 万位以下の低頻度語彙は検討する必要性がないと判断した。そして，この 5 万語に対して散布度と単語親密度を付与した。散布度の付与については 4.2.2.で，単語親密度の付与については 4.2.4.で詳しく述べる。

表 41　再構築したコーパスの総語数（未知語・空白・記号等を除く）

	総語数（万語）
延べ語数	9463
異なり語数	19

その後，5 万語に絞った頻度リストの見出し語について，本表には入れない品詞（助詞，助動詞，接辞）と，分野語（数詞，指示代名詞，曜日など）を削除した。このように，本表に含めない品詞，分野語を除いた結果，本表には約 3 万 6 千語（35938 語）が残った。

4.2.2. 散布度の付与

散布度（DP）は次のように計算した。まず，コーパスのテキストを媒体ごとに全て結合した後，2 万語区切りのテキストに分割し直した。そして，この 2 万語区切りのテキストから，媒体ごとに散布度（DP）を計算し，次に，その媒体ごとの散布度（DP）の平均値を出して，それを全体の散布度（DP）として一次データに付与した。媒体ごとの総語数と，2 万語区切りのテキスト数（ファイル数）は表 42 の通りである。

表 42　2 万語に分割したテキストの数

媒体	総語数（万語）	2 万語区切りファイル数
書籍（PB）	2855	1428
雑誌（PM）	444	222
新聞（PN）	137	69
書籍（LB）	3038	1519
白書（OW）	488	244
検定教科書（OT）	93	47
広報紙（OP）	376	188
ベストセラー（OB）	374	187
Yahoo!知恵袋（OC）	1026	513
Yahoo!ブログ（OY）	1019	510
韻文（OV）	25	13
法律（OL）	108	54
国会会議録（OM）	510	255

より正確な散布度情報を得るため，媒体ごとに可能な限り多くのテキストをサンプリングしたいが，計算上，ファイルはすべての媒体から同数サンプリングしなければならない。ファイル数は最も大きな媒体は流通書籍（LB）で 1428 ファイルを確保できるが，最も小さな韻文（OV）では 13 ファイルしかない。

しかし，韻文（OV）は特殊なテキストジャンルで，日本語教育を目的とした場合，主要なものではない。本研究の目的においては，全体の散布度の正確性を高めるほうが優先されるので，ここでは韻文（OV）を散布度の計算から除外した。

韻文（OV）を除外すると，それ以外の媒体からは，ほぼ 50 ファイルずつサンプリングが可能である[20]。また，新聞（PN）を除外すれば，約 100 ファイルのサンプリングも可能である[21]。しかし，新聞は日本語教育においてもよく利用される媒体であり，韻文（OV）のように除外すべきではないと考えた。

そこで，2 万語区切りのテキストを 50 ファイルサンプリングした場合（総語数 100 万語）と，100 ファイルサンプリングした場合（総語数 200 万語）では，どの程度 DP の値に違いが出るかを検証した。ここでは，最もサイズの大きい媒体である書籍

[20] OT（教科書）は 47 ファイルである。
[21] 法律（OL）は再構築したコーパスに含めないので，ここでは考慮外とした。

（LB）を使った。その結果が表 43 および図 4 である[22]。

表 43　頻度順に見た 100 万語と 200 万語の DP 平均比較（LB:書籍）

頻度順位	200 万語サンプリング DP 平均	100 万語サンプリング DP 平均
1-1000	0.34	0.42
1001-2000	0.42	0.51
2001-3000	0.47	0.57
3001-4000	0.52	0.62
4001-5000	0.56	0.66
5001-6000	0.58	0.69
6001-7000	0.61	0.72
7001-8000	0.63	0.74
8001-9000	0.64	0.76
9001-10000	0.66	0.78

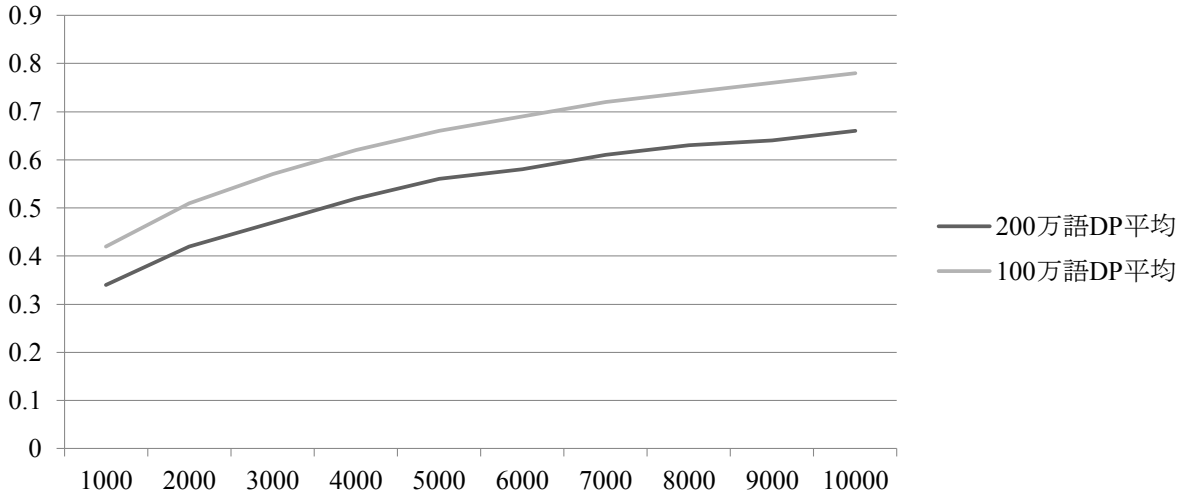

図 4　頻度順に見た 100 万語と 200 万語の DP 平均比較（LB:書籍）

DP の平均値は頻度順位 1000 位ごとに見ていった。頻度ランク 1～1000 までの範囲を見ると，100 万語でサンプリングした場合の DP 平均値は 0.34，200 万語の場合は 0.42 であった。また，4001～5000 の範囲では，100 万語サンプリングが 0.56，200 万語が 0.66，9001～10000 の範囲では，100 万語サンプリングが 0.66，200 万語が 0.78 となっていた。このように，頻度ランクが上がるにつれて，100 万語と 200 万語

[22] DP は 0～1 の範囲で示され，0 の場合が完全に均一に分布している状態を，1 に近づくほど分布に偏りがある状態を示す。

の場合では若干差が開いてくるが，頻度順位 1 万語までの範囲であれば，100 万語サンプリングと 200 万語サンプリングの差は，ほぼ 0.1 前後の範囲で収まり，大きな違いは出なかった。

このように検討した結果，本研究では各媒体から 100 万語（2 万語×50 ファイル）をサンプリングすることにした[23]。表 44 は媒体ごとの DP の平均値を示す。

この結果を見ると，語彙分布の安定度は媒体別にも若干の違いはありそうである。新聞（PN）や Yahoo!ブログ（OY）などは他の媒体と比べて全体的に語彙分布が安定しているが，検定教科書（OT）は高頻度の部分から比較的不安定になる傾向が見られる，また，白書（OW）や国会会議録（OM）は，高頻度部分の DP は平均的であるものの，低頻度の部分の DP の上昇率がやや高く，分布が安定しない語が多くなる傾向があることがわかる。このように媒体によって語彙分布傾向はやや異なるが，頻度レベルから見た DP の上昇率に大きな違いは見られなかった。

表 44　頻度順に見た 100 万語の DP 平均比較

頻度ランク	書籍 LB	書籍 PB	新聞 PN	雑誌 PM	ベストOB	知恵袋 OC	会議録 OM	教科書 OT	広報誌 OP	ブログ OY	白書 OW
1-1000	0.42	0.45	0.29	0.38	0.37	0.39	0.43	0.52	0.36	0.27	0.42
1001-2000	0.50	0.54	0.37	0.47	0.46	0.48	0.55	0.59	0.45	0.36	0.52
2001-3000	0.57	0.59	0.43	0.53	0.53	0.55	0.64	0.65	0.53	0.44	0.60
3001-4000	0.62	0.64	0.49	0.58	0.59	0.61	0.70	0.69	0.59	0.51	0.66
4001-5000	0.66	0.68	0.54	0.62	0.64	0.66	0.74	0.72	0.64	0.56	0.71
5001-6000	0.70	0.72	0.59	0.66	0.67	0.69	0.77	0.75	0.68	0.61	0.75
6001-7000	0.72	0.74	0.62	0.69	0.70	0.72	0.80	0.77	0.72	0.64	0.77
7001-8000	0.74	0.76	0.65	0.71	0.73	0.75	0.82	0.79	0.74	0.67	0.80
8001-9000	0.76	0.78	0.68	0.73	0.75	0.77	0.83	0.81	0.76	0.70	0.82
9001-10000	0.78	0.80	0.70	0.75	0.76	0.79	0.85	0.82	0.78	0.72	0.83

このように本研究では，まず，各媒体 100 万語ずつサンプリングしたコーパスから語彙の散布度（DP）を媒体別に計算し，その 11 媒体間の平均値を出した。そして，これを全体の DP としてデータに付与した。

[23] OT のみ 47 ファイル

4.2.3. 有用度の算出と付与

本研究では，4.2.2.で計算した散布度（DP）の逆数に頻度の対数を掛けた値を語の有用度指標として利用した（3.3.3.参照）。計算は Excel の関数を利用して行い，データに付与した。

4.2.4. 単語親密度の付与

ここでは単語親密度情報の付与について説明する。本研究では可能な限り客観的に語彙を選定することを目指している。そのため，語彙はコーパスの出現頻度と統計指標に基づいて重要度を定め，ランキングしていく。しかし，そのままでは単なる頻度表のようなものになり，本研究で目指す語彙表としては不十分である。そこで本研究では，日本語教育的観点から語彙の難易度を見直すため，単語親密度を用いて重要度ランク順のリストを再配列するという作業を行った。その手順は以下の通りである。

まず，Excel の VLOOKUP 関数を利用して，データに天野・近藤（1999）の文字音声親密度を付けた。単語親密度には，文字音声親密度，音声親密度，文字親密度の 3 種類があるが，本研究では日本語教育で一般的に使われている文字音声親密度を利用した。本研究で作成する語彙表は書き言葉コーパスを使用しており，主に書き言葉の理解語彙を研究対象としているが，実際の教育現場では教科書などの書き言葉の素材も音声化して読まれ，特に，初級段階では教科書中の会話文などに多く触れることになる。したがって，ここでは文字音声親密度を使うのが適当であると判断した。

ただし，「茶まめ」で形態素解析した短単位の語彙素と，天野・近藤（1999）の語彙リストの見出し語では，単位や表記が完全に一致しない場合もある。これらについては自動で値を付与することができないので，一つ一つ目で見て修正作業を行った。一致しないパターンには以下のようなものがある。なお，BCCWJ は「B」，天野・近藤（1999）は「N」で示す。:

(1) 漢字表記違い：

例）B 淀む・N 澱む，B 宛がう・N あてがう，B 御八つ・N お八つ

これらは表記が違っても意味的に同じものを指すと判断した場合，その単語親密度を付与した。

(2) 語形違い：

例）B すっ・N すっと，B 信ずる・N 信じる

茶まめでは，「すっと」は「すっ＋と」と解析され，「信じる」は語彙素では「信ずる」に変換される。したがって，これらは同一のものを指すため，単語親密度もその値を付与した。

(3) 派生形

例）～め：短め（短い）／大きめ（大きい）

申し込み（申し込む）

仕上げる（仕上げ）

早く（早い）

このような派生形には単語親密度のデータが存在しないものも多かった。そこで，これに関してはその派生形の元の形となる語の単語親密度を付与した。例えば，「短め」には「短い」の値を付与した。

このほか，漢字表記が一致しても，読み方が異なり，実際は別語であるものもあった。例えば，「打ちのめす（うちのめす／ぶちのめす）」「一筆（ひとふで／いっぴつ）」，「聖（セイ／ヒジリ）」，「菖蒲（ショウブ／アヤメ）」などであるが，これらは漢字での表記が一致するため，VLOOKUP 関数を使って機械的に処理すると同じ値が付けられてしまう。そのため，このような語彙は一つ一つ目で見て，手作業で単語親密度を

修正した。

以上のような基準で，単語親密度の修正作業を行った。このように，語彙リストの項目が単語親密度の見出し語表記と完全一致しないものについても，内容的に同じものを指しているか，近いものを指していると判断したものには積極的にその値を付与した。本研究で語彙リストに単語親密度を付与する目的は，語彙レベルを判定する材料の一つとすることであり，単語親密度を厳密にリスト化することではない。したがって，多少の修正を加えても，可能な限り多くの語についてレベル判定の手がかりを得ることのほうを優先した。

また，本研究では，原則的に文字音声親密度を参考に，語彙のレベル分けを行うが，表記の影響によって極端に文字音声親密度が低くなっていると考えられる語彙については，音声単語親密度を参考にした（3.4 節参照）。

4.2.5. 削除した語彙

ここでは，語彙表を作成するにあたり，一次データから削除した語彙について述べる。削除したのは，茶まめで形態素解析した際に付けられる品詞の助詞，助動詞，接頭辞，接尾辞，固有名詞，数詞，指示代名詞，接辞的形状詞，さらに，時に関わる語彙である（表 45）。以下，削除対象とした理由についてそれぞれ説明する。

まず，機能語である助詞と助動詞については削除した。本研究で作成する語彙表は，内容語のリストである。また，日本語教育において，助詞と助動詞は，語彙というよりも文型として扱われることが多い。そのためこれらを削除対象とした。

同様に，接頭辞と接尾辞についても，内容語のリストを作るという目的と合わないため，削除した。接頭辞，接尾辞は，語形成要素として語彙的にも重要な項目であるが，語ではなく形態素である。

表 45　削除した項目

項目	例
助詞	の，に，て，は，を，が，と，で，も，の，から，が，か，か，と，や，ば，の，から，など，よ，ね，まで，へ，て，だけ，って，たり，ながら，な，より，けれど，し，ほど，くらい，の…
助動詞	だ，た，ます，です，ない，れる，ず，られる，てる，せる，たい，べし，り，らしい，ちゃう，つう，や，させる，てる，き，ごとし，まい，じゃ，む，とく，ず，たがる，とる，てく…
接頭辞	御，第，御，大，不，約，小，新，各，無，再，全，中，非，高，総，諸，相，超，副，最，低，大，未，本，被，長，両，現，要，多，今，反，小，短，好，計，翌，女，真，原，下…
接尾辞	的，さん，者，達，つ，等，さ，性，人，人，化，方，等，日，歳，中，様，所，後，家，ちゃん，目，内，上，力，用，生，氏，書，君，館，車，易い，国，型，員，長，学，系，権，物…
固有名詞	日本，トウキョウ，中国，昭和，平成，オオサカ，明治，韓国，エド，キョウト，米国，朝鮮，ホッカイドウ，フジ，ソ連，ヨコハマ，タナカ，ノブナガ，サトウ，フクオカ，自民，ヒロシ…
数詞	一，二，三，四，五，六，十，百，千，万，億…
指示代名詞	ああ，あそこ，あちら，あの，あれ，こう，ここ，こちら，この，これ，そう，そこ，そちら，その，それ，どう，どこ，どちら，どの，どれ，こんな，どんな，あんな，
時に関わる語彙	昼，夕方，晩，夕食，昼食，晩御飯，朝御飯，昼御飯，朝， 日曜，土曜，金曜，月曜，水曜，木曜，火曜，曜日，土日曜， 何曜
接尾辞的形状詞	よう，そう，みたい

地名，人名などの固有名詞は，学習者が住んでいる地域が日本国内か，国外かなどの条件によっても重要度が変わってくる。また，トピックに依存する度合いも高い。そのため,日本語教育基本語彙として他の品詞と同様に重要度や難易度を定めにくい。したがって，今回のリストからは固有名詞を除外した。

数詞と時に関わる語についても削除対象とした。日本語教育では，これらはいずれも初級段階で学習する項目であり，それぞれの重要度の違いを示すことはあまり意味がない。また，頻度や分布に基づくランキングによって，対概念の語彙が別レベルに配置されることは，本研究で作成する語彙表の目的には合わない。例えば,「一」は高頻度で重要だが「十九」は低頻度で上級レベルになったり,「日曜」と「水曜」が異なる語彙レベルになったりする可能性がある。語彙表に語彙の頻度や分布を正確に反映したい場合はそれも有益な情報であるが，本研究で作成する語彙表の目的や日本語教育的観点においては意味のない情報である。

指示代名詞についても，日本語教育では初級文型として扱われる項目で，「ここ・そこ・あそこ・どこ」などは同時に提出されることが普通であるため，ここでは削除対象とした。また，接尾辞的形状詞についても，語彙というよりは文型として扱われる項目であるため削除した。

以上のような理由で，上記の項目はデータから削除し，本研究の語彙表に残らないようにした。これは，これらが日本語学習上重要ではないということではない。これらを語彙リストに入れるためには，学習者はどの教科書で学び文型をどの順で学習するか，対象学習者にとって身近な固有名詞は何か，など多様な観点からの配慮が必要である。しかし，これは本研究で作成する語彙表の目的や方法論とは合わないため，本研究では扱わなかった。

4.3節　語彙のランキングとレベル分け

4.3.では，本語彙表における語彙のレベル認定を行う基礎作業として，4.2.で作成したデータの分析を行う。そして，その分析に基づく語彙の選定とレベル分けに関して説明する。

4.3.1.では，語彙の出現頻度と分布状況から，4.2.で作成したデータの特徴を分析する。ここから,本語彙表における語彙のレベル設定や各レベルの語数設定を検討する。また，これは，頻度ランク何語までが分布の安定した高頻度語彙にあたり，語彙表作成において日本語教育的観点からのランク調整を必要とするのかを考える際にも必要な資料となる。

そこで，まず，頻度順位を 1 万語までを目安とし，1000 語区切りの単位で累積頻度と DP 値の散らばりを見る。さらに，1000 語区切りの各グループの頻度と DP の記述統計をもとにクラスタ分析を行う。本語彙表のレベル分けの際の語数の決定はこのクラスタ分析に基づくグルーピングの結果を参考に行う。

4.3.2.では，4.3.1.で出現頻度と語彙分布の傾向に基づき頻度別に語彙をグルーピングした結果を踏まえ，本語彙表の各レベルを何語区切りにするかという問題について検討する。

4.3.3.では，語彙の選定方法について具体的に述べる。まず，出現頻度と散布度から有用度を算出し，これをもとに各レベルに入れる候補語を選定する方法について説明する。次に，この候補語を単語親密度を基準に再配列し，各レベルの語数を絞り込む過程と，日本語教育における語彙の難易度に対する考え方とは合わない単語親密度特有の傾向が,本語彙表のランク調整に影響しないよう補正を行う点について述べる。また，一部の高頻度語彙について例外的に考慮した点についても記す。最後に，このようにな手順で完成させた本語彙表各レベルの有用度と単語親密度の傾向をまとめ，日本語教育語彙表としての妥当性について述べる。

4.3.1. 語彙分布の分析

4.3.1.1. 頻度の傾向

本語彙表の語彙を選定するにあたり，コーパス頻度ランク上位 1 万語における分布の傾向を知るため，頻度ランク上位 1 万語までの頻度の最高値，最低値，平均値，標準偏差，累積頻度，累積比率を求めた。高頻度語彙である頻度ランク 1000 語までは詳しく見るために 100 語区切りに，1000 語以降は 1000 語区切りに集計を行った（表 46，表 47，図 6，図 7)。

表 46　頻度ランク 1000 語までの累積頻度

頻度順位	最高値	最低値	平均値	累積頻度	累積比率
1~100	2,282,435	36,997	138,321	13,832,121	35%
101~200	36,955	22,920	29,042	16,736,328	42%
201~300	22,896	17,302	19,628	18,699,125	47%
301~400	17,259	13,855	15,429	20,242,001	51%
401~500	13,842	11,771	12,712	21,513,186	54%
501~600	11,768	10,236	11,023	22,615,524	57%
601~700	10,219	9,066	9,611	23,576,598	59%
701~800	9,056	8,156	8,562	24,432,830	61%
801~900	8,141	7,418	7,772	25,209,983	63%
901~1000	7,411	6,752	7,078	25,917,817	65%

表 47　頻度ランク 1 万語までの累積頻度

頻度順位	最高値	最低値	平均値	累積頻度	累積比率
1~1000	2,282,435	6,752	25,918	25,917,817	65%
1001~2000	6,738	3,489	4,801	30,718,493	77%
2001~3000	3,487	2,136	2,726	33,444,040	84%
3001~4000	2,136	1,491	1,786	35,230,309	88%
4001~5000	1,489	1,111	1,286	36,516,482	92%
5001~6000	1,111	873	985	37,501,859	94%
6001~7000	873	703	784	38,285,547	96%
7001~8000	703	577	636	38,921,707	98%
8001~9000	577	486	529	39,450,668	99%
9001~10000	486	416	449	39,899,286	100%

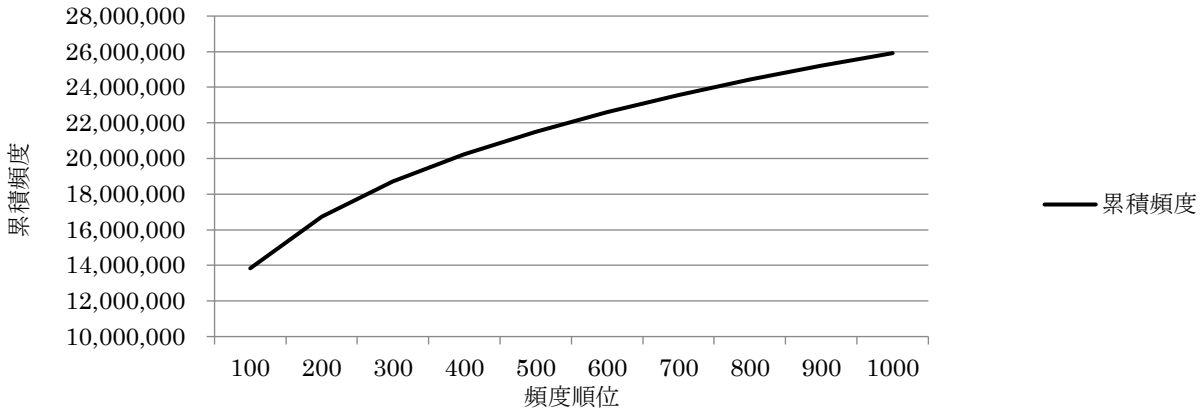

図 5　頻度ランク 1～1000 語までの累積頻度

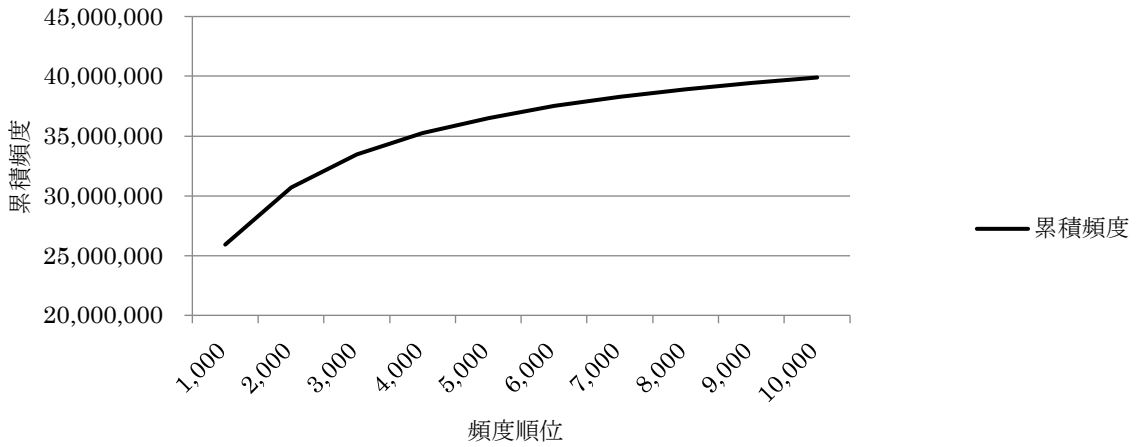

図 6　頻度ランク 1～1 万語までの累積頻度

表 47 の頻度の最高値，最低値，標準偏差を見ていくと，高頻度語彙の部分ほど最高値と最低値の差が大きい。例えば,頻度順位 1～1000 までの頻度最高値は 2,282,435,最低値は 36,997 で，差は 2,245,438 ある。一方，低頻度語彙の部分となる上位 9001～10000 では，最高値が 486，最低値が 416 で，その差は 70 しかない。平均値を見ても，頻度順位が 1～1000 と 1001～2000 の間には大きな差があるが，それ以降の差は急激に小さくなっていく。

累積頻度と累積比率を見ると,上位 1 万位までの累積頻度は 39,899,286 であるが,

頻度上位 100 位までの累積頻度が 13,832, 121 となり，全体の 35%を占めた。このことから，頻度上位 100 位までがコーパス中で最も大きな部分を占める中心的な語群と考えられる（表 46，図 6)。これはジップの法則[24]（Zipf's law: Zipf, 1935, 1949）として知られている。

また，1000 位までで 65%，2000 位までで 77%，3000 位までで 84%と比較的大きく上昇し，4000 位までで 88%，5000 位までで 92%となり，5000 位以降で全体の 90%を超える結果となった。図 7 にも見られるように，頻度上位 3000 位周辺までは累積頻度が大きく上昇するが，それ以降の上昇率は，特に 5000 位以上の部分では，1～2%ずつしか上がらず緩やかになる。

このように，上位 1 万語までの頻度分布を見ると，頻度上位 5000 語までではほとんどの部分を占める。中でも頻度上位 1000 語までの語群は最も大きなグループで，累積頻度の過半数がここに集中している。また，その 1000 語の中でも上位 100 語がコーパスの中核的な語群であることが分かる。

したがって，本研究で作成する日本語教育語彙表のレベル分けを行う際には，この頻度上位 1000 語～2000 語までの語彙に関して特に配慮をすべきである。そして特に，頻度上位 100 語までの語彙については，語彙選定とレベル分けにおいてそれ以下の語彙群とは異なる基準を設定するのも有効な方法であろう。

4.3.1.2. 散布度の傾向

4.3.1.2.では，本研究で算出した散布度（DP）の傾向について調査した。4.3.1.1.で頻度集計した方法と同様，頻度ランク 1～1 万語までを 1000 語区切りにし，その範囲の DP の最高値，最低値，平均値，標準偏差を示したのが表 48 である。

[24] ジップの法則とは，出現頻度が k 番目に大きい要素が全体に占める割合が 1/k に比例するという経験則である。

表 48　1 万語までを 1000 語区切りにした場合の DP の値

頻度順位	最高値	最低値	平均値	標準偏差
1~1000	0.899	0.090	0.506	0.12
1001~2000	0.874	0.438	0.653	0.08
2001~3000	0.943	0.536	0.755	0.07
3001~4000	0.966	0.655	0.822	0.05
4001~5000	0.977	0.700	0.866	0.04
5001~6000	0.991	0.761	0.893	0.04
6001~7000	0.991	0.789	0.912	0.03
7001~8000	0.995	0.820	0.928	0.03
8001~9000	0.995	0.848	0.939	0.02
9001~10000	0.995	0.852	0.947	0.02

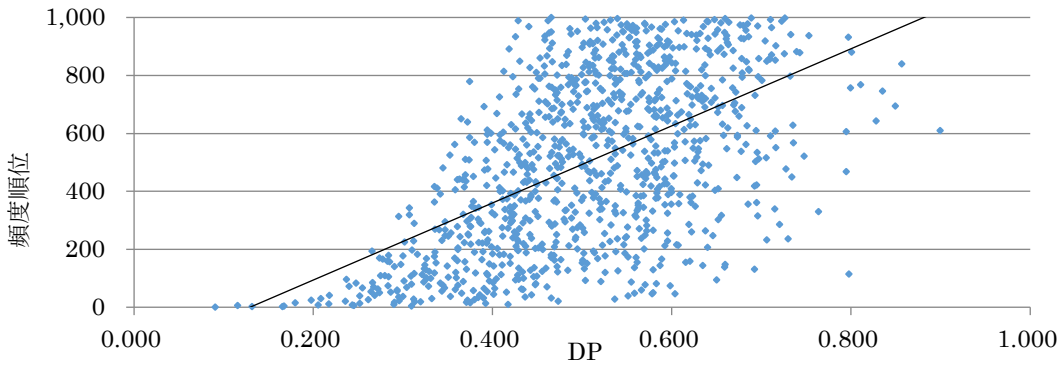

図 7　散布度（DP）値の分布（頻度上位 1~1000 まで）

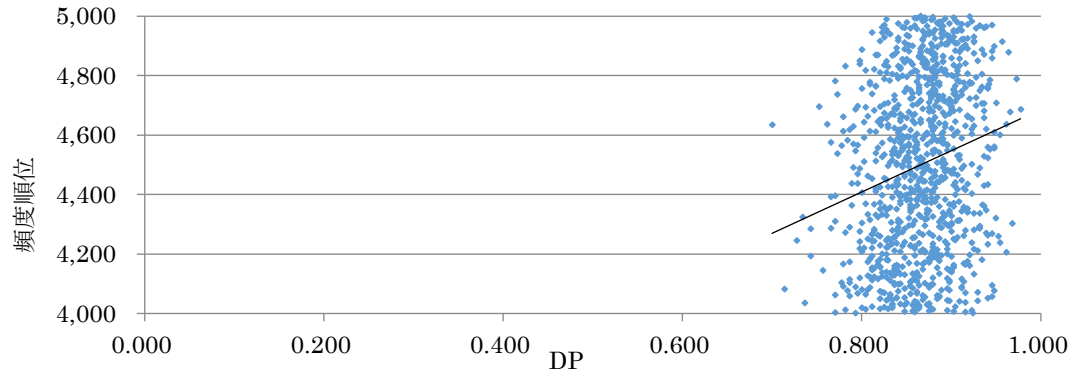

図 8　散布度（DP）値の分布（頻度上位 4001~5000 まで）

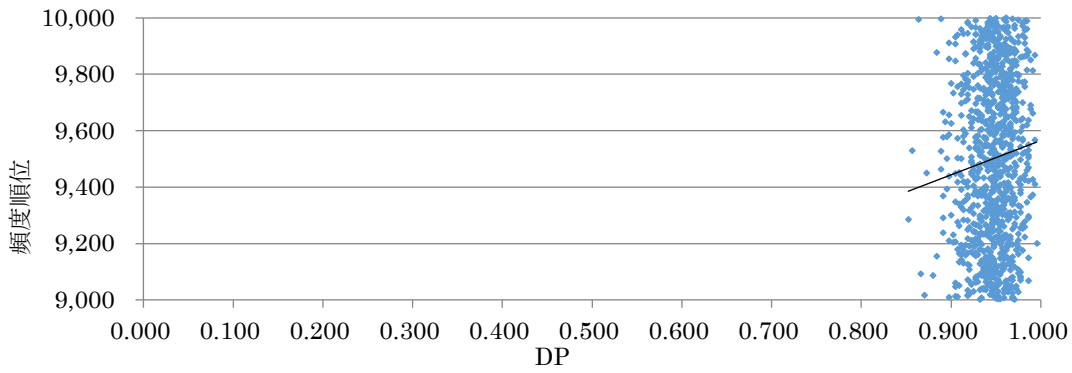

図 9　散布度（DP）値の分布（頻度上位 9001~10000 まで）

表 48 を見ると，DP の平均値は，頻度ランク 1～1000 が 0.506，1001～2000 が 0.653，3000，2001～3000 が 0.755，3001～4000 が 0.822，4001～5000 が 0.866，5001～6000 が 0.893，6001 以降は 0.9 以上へと上昇していく。ここから，語彙分布が比較的安定しているのは頻度ランク 1000～2000 程度の範囲までであると言える。

また，DP 値の分布状況は，図 8，図 9，図 10 の散布図に示した。頻度順位 1～1000 までは，DP の最低値が 0.090，最高値が 0.899 と大きな幅があり，DP の値も散らばっているが（図 8），頻度ランクが下がるにつれてこの差も縮小していき，頻度順位 4001～5000 の範囲では，最低値が 0.655，最高値が 0.966 で，実質的に 0.8 以上のものがほとんどとなる（図 9）。さらに，頻度順位 9001～10000 の範囲になると，DP の値はほとんどが 0.9 以上の部分に密集する（図 10）。このように，頻度順位 4000 あたりからそれ以降は，そのほとんどが分布の安定しない語群であると考えられる。

したがって，本研究で作成する日本語教育語彙表のレベル分けを考える場合，高頻度の部分では頻度にも DP にも幅があり，レベルによって比較的大きな差が生じることが想定できるが，低頻度の部分では，頻度も DP の範囲にも幅がなく近い値に集中するので，レベルごとにあまり差が出ないことが予想される。

なお，表 48 および図 8 にも示されているように，すでに頻度上位 1～1000 の段階から DP の高い語彙，すなわち分布の安定しない語彙が混入してくる。低頻度語に分

布の安定しない語が集中することは理解しやすいが，高頻度語で分布の安定しない語彙が含まれていることには二つの理由が考えられる。一つは，BCCWJ が様々なジャンルのテキストを含むコーパスであることである。特定のジャンルで高頻度でも，別のジャンルでは全くそうではないという場合もある。また，もう一つの理由は，本研究の語彙表が，機能語等を除いた内容語のみで構成されることである（第 3 章参照）。機能語は一般的に高頻度で分布も安定しているが，ここでは機能語を除いた内容語のみで集計しているので，このような傾向がより強く表れていると考えられる。

4.3.1.3. 頻度と散布度の傾向によるグルーピング

4.3.1.2.の調査の結果では，頻度ランク順に 1000 区切りで見た場合，高頻度語彙の部分ほど頻度差が大きく，低頻度語彙の部分ほど小さくなることがわかった。また，散布度（DP）では，高頻度語彙のグループではその値の散らばりに幅があり，低頻度語彙になるほど分布の安定しない語彙が集中する傾向がある。

この傾向をより明確に把握するため，頻度ランク順で 1000 区切りにした 1～1 万語までの 10 のグループについて，頻度，DP の最高値，最低値，平均値，標準偏差（表 49）を変数としてクラスタ分析し（図 10），傾向別にグルーピングを行った（表 50）。

表 49　頻度および　散布度（DP）の記述統計（1000 語区切り）

		1000	2000	3000	4000	5000	6000	7000	8000	9000	10000
頻度	最高値	2282435	6738	3487	2136	1489	1111	873	703	577	486
	最低値	6752	3489	2136	1491	1111	873	703	577	486	416
	平均値	25918	4801	2726	1787	1287	985	784	636	529	449
	標準偏差	93751.7	900.9	392.1	185.2	108.2	68.0	48.5	36.0	25.6	19.90
DP	最高値	0.90	0.87	0.94	0.97	0.98	0.99	0.99	0.995	0.995	0.995
	最低値	0.09	0.44	0.54	0.65	0.70	0.76	0.79	0.82	0.85	0.85
	平均値	0.51	0.65	0.75	0.82	0.87	0.89	0.91	0.93	0.94	0.95
	標準偏差	0.12	0.08	0.07	0.05	0.04	0.04	0.03	0.03	0.02	0.02

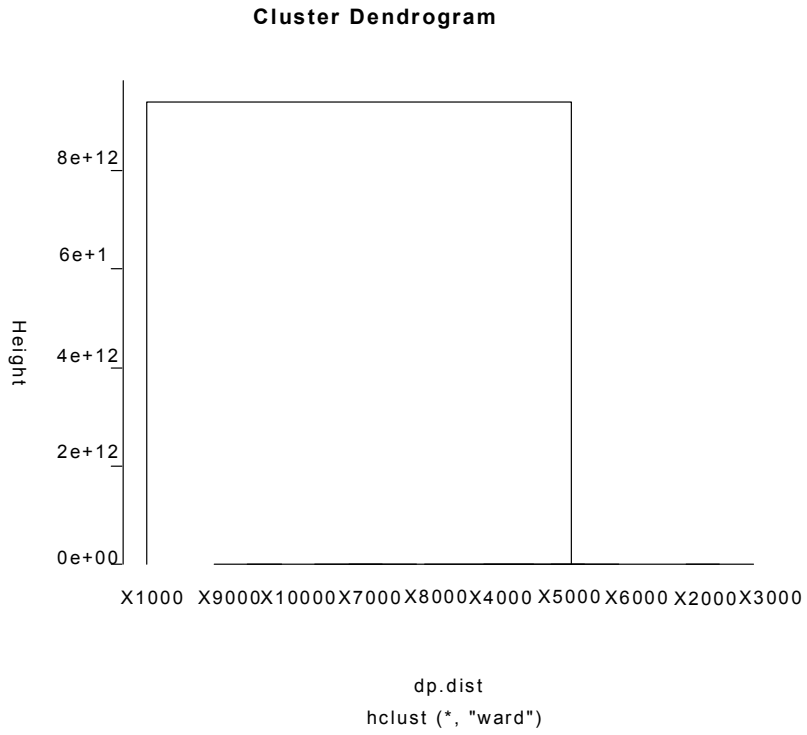

図 10　クラスタ分析の結果（1）

まず，頻度ランク 1000～1 万までを 1000 語区切りにした 10 グループをクラスタリングしたところ，図 10 のような結果となった。これを見ると 1～1000 のグループとその他に二分されている。この結果は，頻度ランク 1～1000 のグループがそれ以外と非常に大きく異なるということを示している。しかし，これだけでは頻度ランク 1001 以下の違いが認識できない。

そこで，次は 1～1000 の部分を除き，1001 以下の 9 グループでのクラスタリングを行った。その結果が図 11 である。

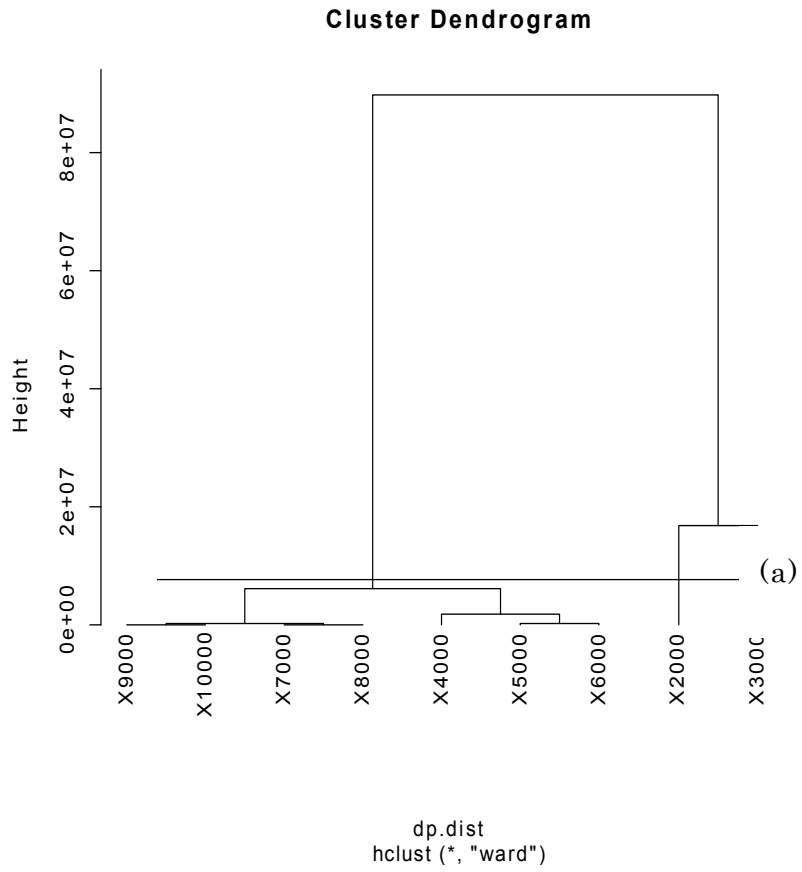

図 11　クラスタ分析の結果（2）

表 50　頻度と分布の傾向によるグルーピング

大分類	中分類	小分類	頻度順位	頻度と分布の傾向
グループ 1 （中核的な語彙）	(1)	①	1～1000	高頻度語 分布が安定した語
グループ 2 （その他の語彙）	(2)	②	1001～2000	高頻度語～中頻度語
		③	2001～3000	中頻度語
	(3)	④	3001～4000 4001～5000・5001～6000	中頻度語～低頻度語
		⑤	6001～10000	低頻度語 分布が安定しない語

このように頻度順位 1～1000 を除いて分析すると，頻度ランク 1001 以下の傾向がわかる。これを見ると，カッティングポイントを(a)とした場合，1001～2000（表中の 2000）と，2001～3000（表中の 3000）と，その他という三分類が可能である。さらに細かく見ていくと，頻度ランク 1001～2000 と，2001～3000 と，3001～6000（表

中の 4000，5000，6000）と，6001～10000（表中の 7000，8000，9000，10000）の四つのグループに分けられる。

これを最初に除外した頻度ランク 1～1000 とともにまとめたのが表 5 である。大分類では，超高頻度語でテキスト中頻度と分布の観点から中核的な部分を占めているグループ 1 と，それ以外のグループ 2 に分けられる。中分類では，1～1000：(1)，1001～3000：(2)，3001～1 万：(3)の 3 グループに分けられる。小分類では，高頻度語で分布の安定した 1～1000：①，高頻度語の 1001～2000：②，高頻度語から中頻度語の範囲で 2001～3000：③，中頻度語から低頻度語にかけた範囲で 3001～6000：④，主に低頻度語彙で分布の安定しない 6001～1 万：⑤のグループにまとめられた。小分類④は，さらに細かく見ると 3001～4000 と 4001～6000 の間に若干の傾向の違いがありそうである。

本研究で作成する日本語教育語彙表では，主観や前例踏襲のような決め方ではなく客観的にレベル分けや語数設定を行う。4.3.1.1 と 4.3.1.2.でも見てきたように，上位 2000 までの語彙は，それ以下の語彙と頻度と DP において異なる傾向が見られ，それはクラスタ分析の結果にも示された。2000 から 5000 までの間は 1000 語区切り，6000 から 1 万の部分では 4000 語で一つのグループとしてもまとめられるが，日本語教育語彙表として使いやすい，実用的な語数設定とは言えない。

語彙のレベル分けは，頻度，散布度だけで決めるのではなく，単語親密度の基準や，日本語教育的観点からの補正も加えて行う。また，実用面についても考慮する。4.3.1.の結果を踏まえつつ，最終的にどのようにレベル分けを行ったかという点については，4.3.2.で詳しく説明する。

4.3.2. レベル分けと語数

4.3.2.では，4.3.1.の調査結果を踏まえつつ，日本語教育的観点からの実用面も総合的に判断し，本研究で作成する日本語教育語彙表のレベル分けと語数設定について検

討した。その結果，表 51 のような分類となった。以下，その根拠について説明する。

表 51　レベル分けの内容

レベル	レベルの内容（語彙表中の順位）	項目数（語数）	累積項目数
2000 語	初級学習者向け（1~2000 番）	2000	2000
4000 語	初・中級学習者向け（2001~4000 番）	2000	4000
6000 語	中級学習者向け（4001～6000 番）	2000	6000
8000 語	中・上級学習者向け（6001~8000 番）	2000	8000
10000 語	上級学習者向け（8001~10000 番）	2000	10000

2000 語レベルは，初級学習者を対象とした 2000 語（1～2000 番）を選定する。2000 という項目数は，4.3.1.でも明らかになったように，そこに選定される語彙が頻度と散布度においてもそれ以下とは異なる傾向を持つ部分であるため，そのように設定した。

また，日本語教育的観点からは，「出題基準」語彙リストの語数も参考にした。旧日本語能力試験では 3 級が初級修了程度とされ，「出題基準」[25]では 1500 語が選定されている。日本語教育では「出題基準」の信頼が厚く，教育や研究目的などで幅広く利用されている。このことから，日本語教育では初級レベルで習得すべき語彙数が 1500 語程度と考えられていると見てよいであろう。ただし，「出題基準」は日本語教育の全般的な指針となることを意図して作られたものではなく，日本語能力試験の試験問題作成という限定的な目的のために作られたものである。また，語彙学習上，初級レベルを修了した学習者の語彙サイズが 1500 語だということに明らかな根拠があるわけではない。したがって，日本語の初級学習者が学習する語彙の数は，目安として 1500 語かその前後であろう。

このようなことを踏まえ，本研究では最初の語数設定を 2000 語とした。2000 語区切りは表 51 の中分類にも合致する。4.3.1.の調査結果によれば，上位 1000 語はコーパス中特に主要な部分を占める語群であり，ここだけを取り出して一つのレベルとして位置付ける方法もある。しかし，日本語教育的観点から検討すると，最初の 1000

25 『日本語能力試験出題基準〔改訂版〕』（2002）独立行政法人国際交流基金・財団法人日本語国際教育支援会，凡人社

語は，初級段階で学ぶ文型や使用する教科書の影響が大きいので，初級教科書に出現する語彙や初級文型に関わる語彙を優先的に取り入れる必要がある。また，本研究で作成する語彙表は日本語の書き言葉を理解するための受容語彙のリストであるが，ごく初級の段階では，学習において受容語彙と産出語彙がほぼ同等に扱われることが少なくない。さらに，日本語教育では 1500 語程度を初級レベルの語彙として見做す前例がある。本語彙表でも一番下のレベルを 2000 語にすれば，ここで初級レベルを完全にカバーした語数を取り出すことができる。このような理由から，本研究では一番下のレベルを 1000 語ではなく 2000 語まで範囲を広げた。

さらに，2000 語レベル以降についても，表 50 の中分類に従い，語彙表の実用面と日本語教育的な語彙の難易度を考慮して語数を設定した。詳しくは以下の通りである。

表 51 の中分類に従えば，1000～2000，3000，4000～6000，6000～10000 の 5 分類になる。しかし，このままの分け方はレベル間で語数が異なりあまり実用的とは言えない。また，レベル分けは，頻度と分布だけで決めるわけではなく，単語親密度の値も基準に入れる。それは以下の方法で行った。

まず，頻度と散布度によって，各レベルに入れる候補語を選定する。その後，単語親密度の高いものから絞り込みを行うという方法を取る。例えば，2000 語レベルの語彙を選定する場合は，頻度と DP の値で上位 3000 語を選び，そこから単語親密度で絞り込みを行って 2000 語を選定する。つまり，頻度と DP で選定を行う段階では，そのレベルよりも 1000 語下までを候補語とする。したがって，4.3.1.3.で行ったクラスタ分析の結果（表 50）では，中分類(2)の範囲全体の中で，2000 語レベルの選定を行うことになる。同様に，その下の 4000 語レベルでは，中分類(3)の範囲の中で，6000 語以下のレベルでは，小分類⑤の範囲の中で語彙を選び，単語親密度による絞り込みを行う。なお，各レベルの語彙選定の方法については，次の 4.3.3.で詳しく述べる。

このような点を考えると，4.3.1.の分析結果を受けても，日本語教育的観点から語彙表の実用面を考えても，本研究作成の語彙表では，2000 語区切りに 1 万語を提示するのが妥当であると判断した。ただし，クラスタ分析の結果では一つのグループにま

とめられる 6000 語から 1 万語までの語彙についても 2000 語区切りで 6000 語，8000 語，10000 語に分けたのは，語彙表の実用面を重視したためである。6000 語レベル以下は，小分類⑤の語彙で，低頻度かつ分布の安定しない語が多い。このような語彙は，本研究では BCCWJ を使って語彙頻度を集計したが，別のコーパスから集計すれば，たとえ同じサンプリング基準で作られたコーパスであっても，その頻度や分布は異なる結果となる可能性もある語彙である。ゆえに，6000 語～1 万語レベルのレベル分けは，2000 語，4000 語とは異なり，主に実用面を重視した便宜的なレベル分けである。

4.3.3. 語彙の選定と日本語教育的観点からのリストの補正

4.3.2.で示したように，本研究で作成する語彙表では 1 万語を 2000 語区切りにした 5 レベルを提示した。ここではその各レベルの語彙を選定するために，統計指標によって定量化した語彙の重要度順リストを日本語教育的観点からどのように補正したかについて述べる。

ここで言う日本語教育的観点からのリストの補正とは，以下の二点を指す。一つは，語彙を統計指標に基づき重要度順にしたリストを単語親密度によって再配列し，さらに語数を絞り込んで各レベルの語彙を確定すること。もう一つは，単語親密度という指標の特徴によって日本語教育的には易しいとされる語彙が難しいレベルに配置されてしまう場合にそれを戻すことである。単語親密度は母語話者がその語について馴染みがあるかどうかを示す指標で，日本語教育では語彙の難易度を定める場合に利用されることもある。ただし，日本語教育的な語彙の難易度と完全に一致するものではないので，本研究では単語親密度の特徴を見極め，必要に応じて補正を行う。

4.3.3.1.では，有用度指標と単語親密度を用いた語彙選定について述べる。一つのレベル（2000 語）を選定するために，まず，リストを有用度順に並べ替え上位 3000 語を取る。そこから，単語親密度によって 2000 語に絞る。

4.3.3.2.では，高頻度語彙の扱いについて述べる。高頻度語彙は 2000 語レベルに集

中する。2000 語レベルの語彙選定に関しては，4000 語レベル以降の語彙を選定する場合とは異なる基準も取り入れる。

最後に，4.3.3.3.では，4.3.3.1.，4.3.3.2.の基準によって語彙を選定した結果の各レベルにおける有用度指標と単語親密度の範囲（最高値，最低値）と平均値について具体的に示し，傾向をまとめる。

4.3.3.1. 有用度指標と単語親密度を用いた語彙選定

4.3.3.1.では，有用度によって各レベルの候補語 3000 語を選出し，さらにそこから単語親密度によって最終的に 2000 語に絞り込みを行った過程について説明する。その手順は以下の通りである。

まず，データ全体を有用度指標順に並べ替え，上位 3000 語を 2000 語レベルに入れる語彙の候補として残した。次に，その 3000 語を単語親密度順に並べ替え，上位 2000 語を 2000 語レベルの語彙として選定した。次の 4000 語レベルでは，前のレベル（2000 語レベル）の語彙選定から落とした語と，残りのデータの語彙を合わせて，再び有用度指標順に並べ替え，候補語の 3000 語を選出し，その後，単語親密度順に並べ替え，上位 2000 語を 4000 語レベルの語彙として選んだ。6000 語レベル以降の語彙も，原則的にこのような手順で語彙選定を行った。これが語彙選定方法の大枠である。

また，上記の方法による語彙のレベル分けが，日本語教育的観点では不適切と判断したものについては補正を行った。それは，表記の問題により，単語親密度の高低と，日本語教育的観点での語彙の難易度が，明らかに一致しない場合である（3.4.参照）。

なお，本語彙表の作成は統計指標を主軸として行っており，単語親密度は主な基準というよりもレベル認定における一つの目安や参考として利用している。これには以下の二点がある。一つは，単語親密度が単語の主観的特性値であることである。日本語教育では，単語親密度が難易度指標の一つとして利用され，単語親密度によって日本語の基本語彙を選定する試みなどもあるが（金杉他，2002），本研究では，可能な

限り客観的な基準によって語彙を選ぶことを目的としている。

もう一つの理由は，先にも述べたような単語親密度の傾向が，日本語教育的観点から見た語彙の難易度と，必ずしも一致しないことである。漢字表記の難しさのみならず，例えば，モーラが短すぎるために語の意味を想起しにくいものや，多義語でどの意味を考えるべきか迷うものなどに関しても，単語親密度が顕著に低いという現象がしばしば起こる。これらには,本研究で語彙表を作成する際の処理上の問題もあれば，単語親密度の調査方法に由来する問題もある。このような理由から，本研究の語彙表作成においては参考値として扱うのが妥当であると考えた。

4.3.3.2. 高頻度語彙を中心とした日本語教育的観点からのランク調整

ここでは，高頻度語彙に関して行った日本語教育的観点からの語彙選定上の補正について記す（3.4.参照）。最初の 2000 語レベルに選定する語彙は初級レベルの語彙であるが，単語親密度の特徴が日本語教育的観点から見た難易度と一致しない部分があるため，そこには補正が必要である。補正が必要なものとしては，表記の難しさから単語親密度が低くなっているものと，初級文型との関わりから初級段階で導入される可能性の高い語彙が単語親密度の特徴によって 2000 語レベルの選定から落とされてしまうものとの，大きく分けて 2 種類がある。

まず，表記の問題について述べる。語彙表の語彙は，原則的に 4.3.3.1.の方法で選定したが，単語親密度の特徴によって，日本語教育的には易しいと考えられる高頻度語彙も，単語親密度の値が低いために選定から落ちるということがしばしばある。例えば，以下のような語彙である。

(1) 「為る（する)」，有用度指標順の順位：1 位，文字音声単語親密度：3.281
「侭（まま)」，有用度指標順の順位 67 位，文字音声単語親密度：3.375
「遣る（やる)」，有用度指標の順位 97 位，文字音声単語親密度：4.719

これらはいずれも頻度上位 100 位以内に入る語彙である。先にも述べたように，頻度上位 100 位までの累計頻度は，1 万語までの累計頻度のうちの 35%を占め，コーパスの主要な部分となる中核的な語彙であり，表記が難しいために 2000 語レベルの選定から落ちることは，本研究の語彙表作成の目的に反している。これらは語そのものの難易度の問題というよりも，単語親密度の調査方法や，形態素解析の語彙素表記などによる処理上の問題と言える。さらに，日本語教育的観点から見ても，上記のような語彙は初級レベルで扱われることが多く，初級文型との関わりの中で不可避なものも多い。そこで，これらが単語親密度の特徴を直接的に受けないようにするため，有用度指標順の上位 100 位までの語彙に関しては最重要語彙として，単語親密度に関わらず 2000 語レベルの語彙に留めた。

さらに，上位 100 位以降でも，初級文型との関わりの中で避けて通ることのできない高頻度語彙もある。高頻度語彙の中には，初級文型との関わりが深い語彙も多く含まれるが，中には単語親密度が低いために 2000 語レベルに入らないものもある：

(2) 「呉れる（くれる)」，有用度指標順位：126 位，文字音声単語親密度：3.438，
音声単語親密度：5.625
「積もり（つもり)」，有用度指標順位：401 位，文字音声単語親密度：5.188
，音声単語親密度：5.094
「いらっしゃる」，有用度指標順位：401 位，文字音声単語親密度：5.562
，音声単語親密度：5.562

「くれる」は「あげる」や「もらう」などとともに初級文型として扱われることの多い項目であるし，「つもり」も同様である。また，「いらっしゃる」のような敬語動詞も，日本語学習においては初級後半で提出されることが多く，2000 語レベルの語彙として扱われるべきである。そこで，このような一般的に初級レベルで扱われる語彙に関しては，単語親密度の高低に関わらず，有用度指標の順位が高ければ，2000 語レ

ベルに残すこととした。

初級文型に関わる語彙については，上記のような補正を行ったが，中級レベル以上のものに関しては，特にこのような手を加えなかった。その理由には，中級以降は教材等が多様化し，初級レベルほど，学習すべき文型や語彙に関してはっきりとしたコンセンサスがないことが挙げられる。そのため，本研究作成の語彙表では，2000 語レベルの中では，高頻度語彙に関して上記のような補正を行ったが，4000 語レベル以上では，原則的に有用度指標と単語親密度の基準のみによってレベル分けを行った。

4.3.3.3. 各レベルの有用度指標と単語親密度の範囲と平均値

本節では，これまで語彙の選定方法について述べてきた。ここからは，そのレベル分けの中で定められた，各レベルにおける有用度指標と単語親密度の範囲（最高値，最低値）と平均値について具体的に示し（表 52，図 13，図 14），傾向をまとめる。

表 52　各レベルの有用度指標と単語親密度の範囲と平均値

レベル	有用度の範囲	有用度の平均値	有用度順位の範囲	単語親密度の閾値	単語親密度の平均値
2000 語	162.14-9.75	16.01	1 位-2880 位	5.844	6.158
4000 語	28.47-7.92	9.87	110 位-4931 位	5.514	5.871
6000 語	26.51-7.05	8.10	138 位-6952 位	5.281	5.681
8000 語	25.86-6.47	7.20	188 位-8968 位	5.031	5.524
10000 語	27.29-6.04	6.58	128 位-11050 位	4.781	5.394

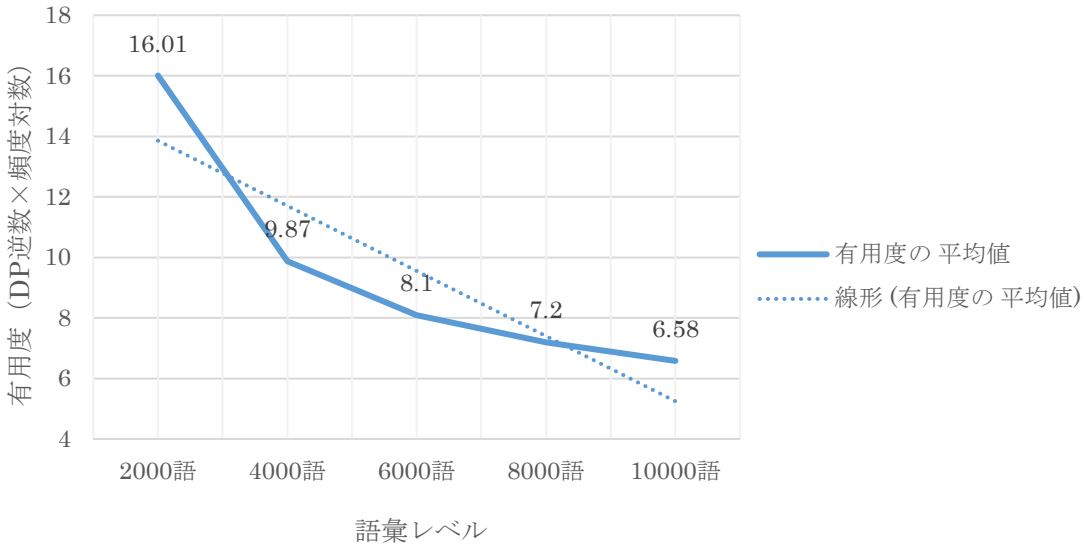

図 12　有用度の平均値

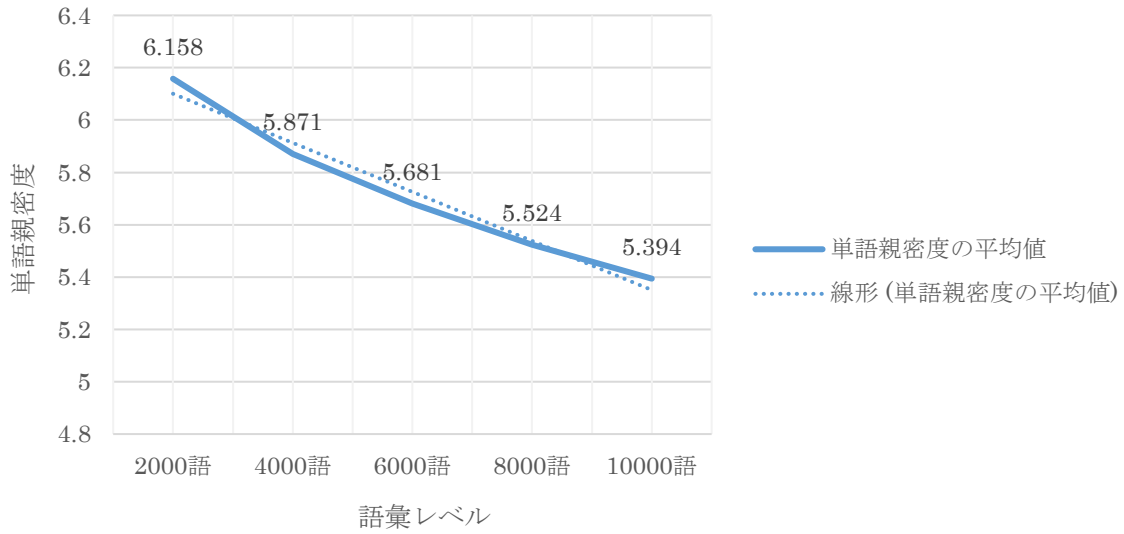

図 13　単語親密度の平均値

まず，平均値を見ると，2000 語レベルに近づくほどレベル間の有用度と単語親密度の平均値の差は大きくなり，1 万語レベルに近づくほど小さくなる傾向がある。その逆の現象は，どのレベル間にも見られなかった。

表 52 の有用度の範囲を見ると，2000 語レベルでは 162.14～9.75 と広範囲だった。

一方，それ以下のレベルでは，4000 語レベルが 28.47～7.92，6000 語レベルが 26.51～7.05，8000 語レベルが 25.86-6.47，1 万語レベルが 27.29-6.04 と，徐々に下がっていくが，レベル間にあまり大きな差はなかった。図 13 にも示されているように，2000 語レベルの有用度の平均値も,それ以下のレベルに比べて際立って高い。これは，2000 語レベルに最重要語とも言える高頻度語彙が集中していることが要因と考えられる。

しかし，単語親密度のほうは有用度ほど極端なレベル間の差はなかった。どのレベル間でも，また，単語親密度の平均値は大まかに見ると単語親密度 2～3 前後の値で推移している（2000 語レベル：6.158，4000 語レベル：5.871，6000 語レベル：5.681，8000 語レベル 5.524，1 万語レベル：5.394）。それは，単語親密度の閾値を見ても同様である（2000 語レベル：5.844，4000 語レベル 5.514，6000 語レベル：5.281，8000 語レベル：5.031，1 万語レベル：4.781）。本語彙表の語彙も 2000 語区切りに等間隔に分けたものなので，レベル間の単語親密度の差に偏りがあまりないことは，語彙表の完成度としても良い結果である。

有用度順位の範囲は，2000 語レベルが 1 位-2880 位，4000 語レベルが 110 位-4931 位，6000 語レベルが 138 位-6952 位，8000 語レベルが 188 位-8968 位，1 万語レベルが 128 位-11050 位，となっている。8000 語レベルや 1 万語レベルのような部分にも 100 位台の高頻度語彙が少しずつ入っている。これは，単語親密度による候補語の絞り込みの際，単語親密度の値が低いために落とされた一部の高頻度語彙が，次のレベルで基準内に入り，そのレベルに位置づけられた結果である。

4.3.4. 4.3.のまとめ

ここでは 4.3.について簡潔にまとめる。4.3.では，語彙のランキングとレベル分けについて述べた。まず，コーパスの語彙分布について分析し，それによってレベル分けと語数設定を決めた。そして，そのレベル分けと語数設定に従い，有用度指標と単

語親密度に基づいて語彙を選定した。また，一部の高頻度語彙に関しては，有用度指標と単語親密度だけでなく,日本語教育的観点からの補正も加えてレベル調整をした。

コーパスの語彙分布は頻度と散布度の傾向を調査した。その結果を踏まえ，語彙のレベル分けは，2000 語レベル，4000 語レベル，6000 語レベル，8000 語レベル，10000 語レベルの 2000 語区切り，5 レベルを設定した。

このレベル分けの枠組みに合わせて，語彙のランキングと選定を行った。語彙のランキングは，散布度に頻度を掛け合わせた有用度指標を基本とし，単語親密度で調整するという方法をとった。具体的には，まず，有用度指標でランキングし，有用度ランク上位 3000 語を,2000 語を選定するための候補語とし,単語親密度順に並べ替え，そこから単語親密度ランク上位 2000 語を残すという方法で行った。

また，一部の高頻度語彙に関しては，日本語教育的観点からの補正が必要と考え，調整を行った。この調整を行ったのは 2000 語レベルの中だけである。上記の方法の場合，有用度が高くても単語親密度が低い語彙は，語彙選定から落とされることになるが，一部の語彙は単語親密度の特徴により低い値となっている。そのため，そのような特徴を受けて低い値となっていると判断した語彙は,2000 語レベルの中に残した。また，日本語学習上初級レベルで学ぶ文型と深く結びついている語彙に関しても，一部，2000 語レベルの中に残した。

このようにして，語彙のランキングとレベル分けを行った。各レベルの有用度の平均値は，2000 語レベルで 16.01，4000 語レベルで 9.87，6000 語レベルで 8.10，8000 語レベルで 7.20，10000 語レベルで 6.58 となった。また，単語親密度の平均値は，2000 語レベルで 6.158，4000 語レベルで 5.871，6000 語レベルで 5.681，8000 語レベルで 5.524，10000 語レベルで 5.394 となった。コーパスの出現頻度や分布の性質上，有用度では，2000 語レベルが突出して高いが，4000 語レベル以下は緩やかに下がっている。また,単語親密度では,レベル間に極端な差や偏りがない結果となった。この結果は，語彙表のレベル分けが，頻度や分布の観点からも，単語親密度の観点からも，バランスよく行われたことを示している。

第5章　語彙表の評価

第 5 章では，本研究作成の語彙表を評価するため，様々なジャンルのテキストにおける本研究作成の語彙表のテキストカバー率を調査する。本研究作成の日本語教育語彙表は，一般の成人日本語学習者が日本語の書き言葉を読んで理解するために必要な語彙を選定することを目的としている。日本語学習の初期段階においては，日本語の教科書や日本語教育用に加工された読解テキストなどを中心に読むことが多い。そして，学習が進むにつれて，新聞や小説・エッセイや Web サイトの情報等の日本語母語話者向けに書かれた一般のテキストを読む機会が増えていく。それは，日本語学習を目的とし，生教材として読む場合もあれば，学習者の興味や関心に基づいて読む場合や，日本国内の学習者は特に，必要な情報を得るために読む場合もある。

そこで，ここでは 5.1.で日本語教育用に加工されたテキストについて，5.2.で一般のテキストについて調査を進める。日本語教育用に加工されたテキストとしては，5.1.1.でレベル分けされた読解テキストとして旧日本語能力試験の読解過去問題を[26]，また，5.1.2.で日本語読解教材用に書き下ろされたテキストとして中上級学習者向けに書かれた小説・エッセイを使う。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 サイトを用いる。また，書き言葉との違いを見るため，話し言葉コーパスも利用する。5.2.1 では利用した一般のテキストの詳細について説明し，5.2.2.ではテキストカバー率を比較する。

5.1節　日本語教育用テキストにおけるテキストカバー率

5.1.では，日本語教育用に加工された読解テキストにおける本研究作成語彙表のテキストカバー率調査を行う。日本語教育用に加工された読解テキストには，日本語能力試験の読解問題において過去に出題された読解テキストと，日本語教材として書き

[26] 日本語能力試験は 2010 年に改訂されたが，新試験の過去問題は公開されていないため，本研究では比較的新しい旧試験の過去問題を利用した。

下ろされた中級レベルの読解用テキストを使用する。さらに，本研究作成の語彙表をより客観的に評価するため，「出題基準」語彙リストとの比較を行う。

5.1.1. 日本語能力試験過去問における本研究作成語彙表のテキストカバー率

まず，旧日本語能力試験の読解問題において過去に出題された読解テキストにおける本研究作成の語彙表のテキストカバー率を調査した。

旧日本語能力試験は高いレベルから順に 1 級，2 級，3 級，4 級の 4 レベルに分かれている[27]が，1.2 級は 5 年分（2005 年～2009 年），3.4 級 10 年分（2000 年～2009 年）を使用した。これは，初級レベルの 3.4 級に用いられる読解テキストの分量が少ないためである。

テキストカバー率調査の方法は，まず，これらのテキストすべてを級別に結合，形態素解析し，語彙リストの形に成形した。そして，本研究作成の語彙表と同じ基準で機能語，固有名詞，対概念語等を削除した。最後に，レベル別に語彙リストとの語の重なりを調査して，テキストにおける本研究作成の語彙表のカバー率を算出した。その結果が表 53 である。

表 53　JLPT 過去問における本研究作成語彙表のテキストカバー率

JLPT 過去問	1 級読解	2 級読解	3 級読解	4 級読解
過去問の総語数（延べ語数） ※（　）は機能語・固有名詞等含む語数	6924 (7154)	5781 (6543)	2682 (3935)	1968 (3362)
1～2000 語レベルでカバーする語数	5119	4579	2385	1722
1～2000 語レベルのカバー率	74%	79%	89%	88%
1～4000 語レベルでカバーする語数	5828	5095	2520	1835
1～4000 語レベルのカバー率	84%	88%	94%	93%
1～6000 語レベルでカバーする語数	6143	5303	2596	1880
1～6000 語レベルのカバー率	89%	92%	97%	96%
1～8000 語レベルでカバーする語数	6335	5451	2609	1892
1～8000 語レベルのカバー率	91%	94%	97%	96%
1～10000 語レベルでカバーする語数	6451	5508	2617	1900
1～10000 語レベルのカバー率	93%	95%	98%	97%

[27] 「出題基準」語彙リストでは，1 級 1 万語，2 級 6000 語，3 級 1500 語，4 級 800 語が示されている。

本研究作成語彙表全体（1 万語リスト）の旧日本語能力試験読解問題テキストカバー率は，1 級～4 級レベルで 93%～98%という結果となった。この結果は，本研究作成の語彙表が日本語能力試験のような日本語教育用に作られたテキストにおいても妥当な語彙選定が行われたことを裏付ける根拠の一つと言える。

また，本研究作成の語彙表のレベル分けについては，旧試験 3 級（1500 語レベル）のテキストで 2000 語レベル語彙リスト（1～2000 語レベル）のテキストカバー率が 89%と若干 90%を下回るものの，旧試験 2 級（6000 語レベル）のテキストで 6000 語レベル語彙リスト（1～6000 語レベル）のテキストカバー率が 92%，旧試験 1 級（1 万語レベル）のテキストで 1 万語レベル語彙リスト（1～10000 語レベル）のテキストカバー率が 93%となっており，該当レベルのテキストカバー率はいずれも 90%前後で，語彙のレベル分けでも日本語教育の実情に合う形になっていると言える。

次に，「出題基準」語彙リストのテキストカバー率についても調査し，本研究作成の語彙表との比較を行った（表 54）。

表 54　JLPT 過去問における「出題基準」語彙リストのテキストカバー率

JLPT 過去問	1 級読解	2 級読解	3 級読解	4 級読解
過去問の総語数（延べ語数） ※（　）は機能語・固有名詞等含む語数	6924 (7154)	5781 (6543)	2682 (3935)	1968 (3362)
4 級リストでカバーする語数	2493	2434	1844	1688
4 級リストのカバー率	36%	42%	69%	86%
3，4 級リストでカバーする語数	3571	3424	2339	1776
3，4 級リストのカバー率	52%	59%	87%	90%
2，3，4 級リストでカバーする語数	5603	4962	2567	1898
2，3，4 級リストのカバー率	81%	86%	96%	96%
1，2，3，4 級リストでカバーする語数	6077	5248	2579	1906
1，2，3，4 級リストのカバー率	88%	91%	96%	97%

「出題基準」語彙リストのテキストカバー率は，旧試験の出題の基準となる語彙リストであるにもかかわらず，本研究作成の語彙表の場合と比べ，いずれのレベルでも低い結果となっている。

これには，二つの理由が考えられる。一つは,「出題基準」語彙リストから得られる総語数が 1 万語ではなく,8000 語程度しか公表されていないためそれを利用するしかないことである。もう一つは，カバー率調査の方法にある。「出題基準」語彙リストの語彙は，本研究作成の語彙表とは語の単位が異なるため，カバー率を比較するためには形態素解析して語の単位をそろえる必要がある。その結果，複合語の一部を拾うことができなかった。また，本研究作成の語彙表と同じ基準で機能語，固有名詞，対概念後等を削除したため，最終的には「出題基準」語彙リスト全体で 7163 語となっている。一方，本研究作成の語彙表は，これらの語彙を削除したうえでの総語数が 1 万語であり,「出題基準」よりも広範囲の語彙が選ばれている。したがって,「出題基準」語彙リストそのものの問題というよりは，テキストカバー率調査の方法が影響している面も少なからずある。

しかし，そうであったとしても,「出題基準」語彙リストの旧試験におけるテキストカバー率は，本研究作成の語彙表よりも高いとは言えない。1 級から 4 級まで，各レベルに対応する読解問題テキストのテキストカバー率を見ていくと，4 級 86%，3 級 87%，2 級 86%，1 級 88%であり，いずれも 90%以下である。

また,「出題基準」のほうが全体で 7163 語のリストであることを踏まえ，2 級の読解テキスト（6000 語レベルを想定）で 1 級レベルリスト（7163 語のリスト）のテキストカバー率を見ると 91%という結果だが，本研究作成の語彙表の 6000 語レベルのリストは 2 級読解テキストを 92%カバーすることができ，リストの語数を単純に比較した場合には，本研究作成の語彙表のほうが 1000 語以上少ないにもかかわらず，テキストカバー率は僅かに上回っている。

したがって，本研究作成の語彙表は,「出題基準」語彙リストとの比較してみても，旧日本語能力試験の読解過去問題におけるテキストカバー率が十分に高い結果となった。このように，過去問のカバー率を見る限り，本研究作成の語彙表は日本語教育語彙表として妥当な語彙選定とレベル分けが行われていると考えられる。

5.1.2. 中・上級用読解テキストにおけるテキストカバー率

次に，試験用読解テキストとは別に，日本語教育用に加工された中・上級用読解テキストにおける本研究作成の語彙表のテキストカバー率を調査した。中・上級用読解テキストとしては，国際交流基金が運営し，日本語教師向けに教材用素材や日本語教育情報を提供している Web サイトの「みんなの教材サイト」[28]に掲載されている「中級読解」と「日本語教育通信エッセイ」を利用した。

「中級読解」は，国際交流基金日本語国際センターが海外司書日本語研修用に制作した『中級読解－日本理解へのステップ－』(国際交流基金日本語国際センター, 1997)から，10 の文章を Web 公開した書下ろし中級学習者向け読解用素材である。「中級読解」テキストの内容の詳細は表 55 の通りである。

表 55 「中級読解」

タイトル	文章の長さ（字数）	文章の種類	レベル
日本人の食生活	800	説明文	中級前半
カルチャーセンター	900	説明文	中級前半
縁起・迷信・占い	900	説明文	中級前半
日本人の大人と漫画	800	説明文	中級前半
日本の女性の結婚観	900	説明文	中級前半
教育－日本の子ども－	1100	説明文	中級前半
ごみの減量化	1100	説明文	中級前半
ラフカディオ・ハーン	1400	伝記	中級後半
日本語の特徴	1400	説明文	中級後半
非言語コミュニケーション	1600	説明文	中級後半

（国際交流基金「みんなの教材サイト」より）

「日本語教育通信エッセイ」は，『日本語教育通信』（国際交流基金）の表紙エッセイから提供されたもので，日本語や日本の社会・文化，外国語学習や異文化などをテーマにした著名人による比較的短いエッセイである。「日本語教育通信エッセイ」の内容の詳細は表 56 の通りである。

[28] https://minnanokyozai.jp/kyozai/mypage/ja/render.do

表 56 「日本語教育通信エッセイ」

タイトル	文章の長さ（字数）	文章の種類	レベル
フリガナのついた本（なだいなだ）	900	エッセイ	中級後半
ことばとリズム（角野栄子）	800	エッセイ	中級後半
戯曲について（鴻上尚史）	1000	エッセイ	中級後半
サッカーから学んだ人生論（川淵三郎）	1000	エッセイ	上級
数のかぞえかた（大岡 信）	1100	エッセイ	上級
日本文化開放の夜明け前（沢 知恵）	1000	エッセイ	上級
おお，しゃれ（阿刀田高）	900	エッセイ	上級
おなじだのに同じでない漫画・まんが・マンガ（高畑 勲）	1000	エッセイ	上級
I know『Shall we ダンス？』（周防正行）	1000	エッセイ	中級後半
日本の夫のジレンマ（土屋賢二）	1000	エッセイ	中級後半
漢字とかな（加藤秀俊）	1100	エッセイ	中級前半
人間は，かぶれる（野田秀樹）	1100	エッセイ	上級
言葉は時代とともに（林真理子）	900	エッセイ	上級
日本語上達の決め手は受身形の使い方にある（呉 善花）	1000	エッセイ	中級後半
思いやり（中西 進）	900	エッセイ	中級後半
やさしく，判りやすい日本語を（福原義春）	800	エッセイ	中級前半

（国際交流基金「みんなの教材サイト」より）

5.1.1.と同様，この 2 種類のテキストすべてを結合，形態素解析し，本研究作成の語彙表と同じ基準で機能語，固有名詞，対概念語等を削除したものと，本研究作成の語彙表との語の重なりを調べる方法で，語彙表のカバー率を調査した。その結果が表 57 である。

表 57 日本語教育用テキストにおける本研究作成語彙表のテキストカバー率

中級読解／日本語教育通信エッセイ	語数とテキストカバー率
過去問の総語数（延べ語数）	6724
1～2000 語レベルでカバーする語数	5045
1～2000 語レベルのカバー率	75%
1～4000 語レベルでカバーする語数	5616
1～4000 語レベルのカバー率	84%
1～6000 語レベルでカバーする語数	5887
1～6000 語レベルのカバー率	88%
1～8000 語レベルでカバーする語数	6042
1～8000 語レベルのカバー率	90%
1～10000 語レベルでカバーする語数	6126
1～10000 語レベルのカバー率	91%

「出題基準」語彙リストの 2 級（中級レベル）が 6000 語，1 級（上級レベル）が 1 万語であることを踏まえると，本研究作成語彙表でも 6000 語レベル以上がこの場合の中・上級の範囲であると考えることができる。

それぞれテキストカバー率を見てみると，6000 語レベルが 88%，8000 語レベルが 90%，10000 語レベルが 91%という結果であった。この結果は，旧試験の 1 級読解問題における本研究作成語彙表のテキストカバー率（6000 語レベルが 89%，8000 語レベルが 91%，10000 語レベルが 93%）とほぼ重なっている。

このことからも，旧試験のテキストカバー率調査の結果と同様，一般の教材として中・上級向けに書かれたテキストにおいても，本研究作成語彙表のテキストカバー率は十分に高いと言える。また，「中級読解」および「日本語教育通信エッセイ」のレベルは「中級前半」から「上級」までという範囲で公表されているが，本調査の結果に照らすと，語彙レベル的には 1 級相当であり，難度の高い語彙も少なからず含まれている可能性が示唆されている。

5.2節　一般テキストにおけるテキストカバー率

5.2.では，日本語教育用に加工されていない一般のテキストで本研究作成語彙表のテキストカバー率を調査する。ここでは一般テキストとして，新聞，小説，Web サイトを用いたのに加え，話し言葉コーパスとの比較も行う。

なお，本研究作成の語彙表は，現代日本語書き言葉の理解を目的としたものだが，ここでは話し言葉についても本研究作成語彙表の特徴を知るための参考として調査を行い，その傾向を見る。また，「出題基準」語彙リストとのテキストカバー率比較も行う。

5.2.1. テキストの選定

テキストカバー率調査に用いる一般のテキストには，成人の日本語学習者が目にする可能性があるテキストであることを基準に，以下の 4 ジャンルを選んだ。その詳細は表 58 の通りである。

表 58　テキストカバー率調査に用いたテキスト

文章の種類	資料名	出版社／作成者	総語数
新聞	『毎日新聞 2010 データ集』	毎日新聞社	1 万語
小説	ポプラビーチ 幻冬舎 Plus Web KADOKAWA	ポプラ社 幻冬舎 角川書店	1 万語
Web サイト	ブログ 「ひらがなタイムズ」 学校ホームページ（大学・高校・中学・小学校） ショッピングサイト Wikipedia	「にほんブログ村」（http://www.blogmura.com/）の人気ブログより 株式会社ヤック企画（http://www.hiraganatimes.com/） 東京大学，他 Amazon，楽天，他 Wikipedia http://www.wikipedia.org/	1 万語
話し言葉コーパス	『BTSJ による日本語話し言葉コーパス』（2011）	東京外国語大学 宇佐美まゆみ監修	1 万語

4 ジャンルのテキストをコーパス化し，テキストカバー率調査用に成形した方法は以下の通りである。まず，コーパスは，原則として一つの作品や記事等から 1000 字ずつランダムサンプリングして作成した。そして，5.1.と同様にテキストを結合，形態素解析した。最後に，本研究作成語彙表と同じ基準で機能語，固有名詞，対概念語等を除いた後，総語数が 1 万語となるようにした。

また，サンプリングの仕方では，同じ作品や Web ページから二回以上サンプリングしないようにした。例えば小説なら，一つの作品から一個所，Web サイトなら一つのブログやホームページなどからは一個所のみをサンプリングした。サンプリングの字数は一回 1000 字ずつを原則としているが，ブログのような Web サイトに関しては，写真等が大部分を占めているため文字の部分が少なく，一つの記事から 1000 字がサ

ンプリングできない場合も一部あった。

このように 1000 字ずつサンプリングし，形態素解析して語の単位にしたものから記号・空白，さらに，機能語，固有名詞等を除くと，1 万語のコーパスを作るのには少なくとも 1000 字×30 ファイル以上はサンプリングする必要があった。このようにテキストカバー率調査用に作成した 1 万語コーパス一つには，様々な出典のものが含まれているので，一つの記事や小説の特徴語が大きく影響するようなことはないと考えられる。

テキストカバー率調査用に，各ジャンルのテキストを 1 万語でそろえた理由は次の二つである。一つは,本研究作成の語彙表が総語数 1 万語の語彙表であることである。1 万語リストのテキストカバー率を見るのに，大規模コーパスのようなものを利用しても,学習者がそれを読んでどの程度語彙を理解できるかという目安にはならないし，また，コーパスサイズが小さすぎても学習者の理解を測る現実的な値は得られない。もう一つは,日本語で A4 サイズの原稿 1 枚分に収められる字数が約 1400 字程度であり，ある程度まとまった分量の書き言葉が読めることの基準としても 1 万語は少ない分量ではなく，特に上級以上の日本語学習者が読む可能性のある書き言葉テキストのカバー率を調査するのに十分な分量であると考えたためである。さらに，4.4.1.で行った旧日本語能力試験 1 級過去問題や，読解教材テキストの総語数も 7000 語弱であることを考えると，これらとの比較もしやすい。したがって，本テキストカバー率調査の目的において，1 万語のテキスト少なすぎず多すぎない適切な分量であると考えた。

テキスト収集の資料については，入手可能であるという実用的な理由で選んだ。新聞では入手可能な比較的新しい新聞コーパスとして,「毎日新聞 2010 データ集」を利用した。また，テキストカバー率調査用には 1 万語という比較的小さなコーパスを作るため，同じ単語が必要以上に重複しないよう，見出しや前文の部分は除き，記事本文だけをサンプリングした。

小説は，Web 上で公開されているものからサンプリングした。「ポプラビーチ」「幻

冬舎 Plus」「Web KADOKAWA」は，各出版社から出版されている書籍の一部を Web 上で閲覧可能にしているもので，最近の作品から近代文学まで様々なジャンルの小説が公開されている。このように，書籍として出版されているものは，個人が Web 上に自由に掲載している Web 小説とは異なり，編集者の手も加えられているし，一般的に読まれる可能性が高いものであり，本研究のテキストカバー率調査にも適していると考えた。

Web サイトコーパスは，ブログ，「ひらがなタイムズ」，学校ホームページ，ウィキペディア，ショッピングサイトからサンプリングした。いずれも，国内外の日本語学習者が閲覧する可能性があることを基準としている。ブログは，「にほんブログ村」（http://www.blogmura.com/）というサイトを利用して，「人気ランキング」上位のものや「注目記事」として取り上げられているものから選んだ。また，ブログはその内容によって旅行，生活，ペット，趣味，スポーツ，仕事，教育，日記等にジャンル分けされているので，サンプリングではジャンルの偏りがないように努めた。「ひらがなタイムズ」は，日本の文化や情報を掲載した雑誌で，学習者向けに英語訳や漢字のルビなどが付いていて，日本語学習者にも適した内容になっている。その記事の一部は Web 上でも公開されており，本研究ではそこからサンプリングを行った。学校ホームページのコーパスは，在籍留学生数が比較的多い大学や，日本語学校，外国人児童生徒がいる可能性のある高校，中学，小学校のホームページからサンプリングした。大学や日本語学校のホームページは留学生や就学生が自ら閲覧する機会があるだろうし，中等教育以下のホームページはその保護者が閲覧する可能性もあると考え，コーパスに加えた。ショッピングサイトは，比較的利用者が多いことを基準に，Amazon（http://www.amazon.co.jp/），楽天市場（http://www.rakuten.co.jp/），価格.com（http://kakaku.com/），Yahoo!オークション（http://auctions.yahoo.co.jp/），楽天オークション（http://auction.rakuten.co.jp/）の五つのサイトからサンプリングを行った。Wikipedia（日本語）では，そのメインページが紹介する「秀逸な記事」（ウィキペディアによって高品質な記事として選ばれたもの）より，ジャンル（哲学・歴

史・社会科学・自然科学・芸術・言語・文学）の偏りがないようにサンプリングした。

話し言葉コーパスは，「BTSJ による多言語話し言葉コーパス」（2011，宇佐美まゆみ監修）を利用した。これは，自然会話分析を目的として収集された会話のトランスクリプトである。本研究では，テキストカバー率調査のために友人同士の会話と，論文指導での会話の二種類に分けて 1 万語ずつコーパスを作った。これは，友人同士の砕けた会話（雑談）と，論文指導のようなアカデミックな場面での語彙使用の違いを見るためである。また，「BTSJ による多言語話し言葉コーパス」は会話分析用に「〈笑い〉」や「《沈黙》」のようなコーディングが含まれているが，本研究では会話部分だけを利用するため，これらはあらかじめ削除した。

5.2.2. 一般テキストにおける本研究作成語彙表のテキストカバー率

ここでは，5.2.1.の方法で収集した一般テキストにおける，本研究作成の語彙表のテキストカバー率を調査した。その結果は表 59 の通りである。

表 59　一般テキストにおける本研究作成語彙表のテキストカバー率

一般テキスト	新聞	Web	小説	話し言葉 雑談	話し言葉 論文指導
一般テキストの総語数（延べ語数）	10000	10000	10000	10000	10000
2000 語レベルでカバーする語数	5495	5816	**6208**	**6488**	**6698**
2000 語レベルのカバー率	55%	58%	**62%**	**65%**	**67%**
4000 語レベルでカバーする語数	6972	7056	7336	7778	7493
4000 語レベルのカバー率	70%	71%	73%	78%	75%
6000 語レベルでカバーする語数	7747	7768	7922	8068	7937
6000 語レベルのカバー率	77%	78%	79%	81%	79%
8000 語レベルでカバーする語数	8258	8278	8271	**8461**	**8520**
8000 語レベルのカバー率	83%	83%	83%	**85%**	**85%**
1 万語レベルでカバーする語数	**8582**	**8543**	**8575**	**8556**	**8726**
1 万語レベルのカバー率	**86%**	**85%**	**86%**	**86%**	**87%**

一般テキストは語彙の難易度が高いため，1 万語レベルを中心に結果を見ていく。

一般テキストにおける本研究作成語彙表の 1 万語レベルのテキストカバー率は 85%～87%の範囲となった。いずれも 90%を超えず，日本語教育用の上級用テキストにおけるカバー率調査の結果と比べると，やや低いように感じる。

しかし，日本語教育用上級用テキストのコーパスが 7000 語規模のものであることや，日本語教育用テキストと語彙をコントロールしていない一般のテキストとの違いを考慮に入れると,このテキストカバー率は決して低くはない値と考えられる。特に，本調査の結果が機能語，固有名詞，対概念語等を抜いた値であることを加味すると，本研究作成語彙表の一般テキストにおけるテキストカバー率は高く，日本語教育用テキストのみならず，一般のテキストにも十分対応できる語彙選択が行われていると言えるのではないだろうか。

テキストジャンル別に見ると，ジャンルによるテキストカバー率にはあまり大きな違いがないと考えられる。ただし，細かく見ていくと，書き言葉よりも話し言葉のほうがややカバー率が高い傾向も伺える。

話し言葉と書き言葉それぞれのテキストのカバー率を 2000 語レベルで見ると，書き言葉である新聞が 55%，Web が 58%，小説が 62%であるのに対し，話し言葉（雑談）が 65%，話し言葉（論文指導）67%という結果になっている。本研究の語彙表は書き言葉の理解を目的としているが，この結果は，やはり話し言葉のほうが語彙のバリエーションが少ないという一般的な傾向を反映しているのかもしれない。

また，BTSJ に多い学生同士の雑談などではくだけた表現やスラング的な語彙も豊富に用いられることが予想されるが，論文指導のようなやや改まった場面ではそのような影響が少ないためか，最もテキストカバー率が高い。これは，本研究語彙リストが書き言葉に対応できていないというよりも,2000 語レベルからでも書き言葉のみならずアカデミックな話し言葉に対応することができる語彙表であると考えられる。

それは，1 万語レベルまで見てみるとより明らかである。2000 語レベルでは新聞が 55%であるのに対し，話し言葉（論文指導）が 67%であり 12%もの差があったが，1 万語レベルで見ると，新聞が 86%，小説が 85%，Web が 86%，話し言葉（雑談）が

86%，話し言葉（論文指導）87%と，その差がほぼ解消している。このように，1 万語レベルの語彙まで含めると，どのテキストジャンルにも対応することができる。

次に，一般テキストにおける，本研究作成語彙表と「出題基準」語彙リストのカバー率を比較した。前にも述べたように,「出題基準」語彙リストの語の単位は形態素解析の短単位とは異なるので,「出題基準」語彙リストを調査用に成形した。まず，形態素解析して単位をそろえ，本研究語彙表と同様に機能語，固有名詞，対概念語等を一部削除した。

その結果，1 級～4 級のすべての語彙を合わせると 7163 語になった。「出題基準」によると，1 級レベルは 1 万語相当とされているが，本調査で比較できる語彙は 7163 語になったため，この 7163 語のカバー率と，本研究語彙表の 1 万語レベルを比較するのではなく，6000 語レベルおよび 8000 語レベルで，テキストカバー率の比較を行った。その結果が表 60 である。

一般テキストにおける「出題基準」1 級リスト（1 級から 4 級まですべて合わせたもの：7163 語）のテキストカバー率は，新聞が 75%，Web が 76%，小説が 80%，話し言葉（雑談）が 84%，話し言葉（論文指導）が 86%であった。

一方，本研究作成の語彙表の 8000 語レベルでは，新聞が 83%，Web が 83%，小説が 83%，話し言葉（雑談）が 85%，話し言葉（論文指導）が 85%であり,「出題基準」1 級リストより話し言葉（論文指導）ではわずかに 1%低かったものの，それ以外ではすべて上回っていた。

さらに, 6000 語レベルと比較してみても, 新聞が 77%, Web が 78%で,「出題基準」1 級リストより 1000 語以上も語数が少ないにもかかわらず,テキストカバー率では本研究作成の語彙表が上回り，さらに，小説，話し言葉（雑談）でもほぼ同程度の値が得られた。

表 60　一般テキストカバー率における「出題基準」との比較

一般テキスト	新聞	Web	小説	話し言葉（雑談）	話し言葉（論文指導）
一般テキストの総語数（延べ語数）	10000	10000	10000	10000	10000
1級リスト（7163語）でカバーする語数	7512	7582	7976	8430	8629
1級リストのカバー率	75%	76%	80%	84%	86%
6000語リストでカバーする語数	7747	7768	7922	8068	7937
6000語リストのカバー率	77%	78%	79%	81%	79%
8000語リストでカバーする語数	8258	8278	8271	8461	8520
8000語リストのカバー率	83%	83%	83%	85%	85%

したがって,「出題基準」語彙リストと本研究作成語彙リストを比較した場合，書き言葉の一般テキストのカバー率では，本研究作成語彙リストが高い値となることが明らかになった。また，話し言葉コーパスでは，特に論文指導のほうで,「出題基準」がやや高めの値となっていた。このことから，本研究作成語彙リストに比べて「出題基準」語彙リストは，論文指導のようなやや改まったアカデミックな会話に使われるような語彙を多く含むという特徴を持っている可能性も示唆された。

第5章では，5.1.で日本語教育用に加工された読解テキスト，5.2.では一般のテキストにおける本研究作成の語彙表のテキストカバー率を調査し，成人日本語学習者の書き言葉の理解を目的とする日本語教育語彙表として適切な語彙が選ばれているかということを評価した。

その結果，日本語教育用テキストでも，一般のテキストでも，本研究作成の語彙表はカバー率が高いことが明らかになった。

また，それは，従来から日本語教育の現場や研究で様々な目的で使われてきた，いわば日本語教育語彙表の「定番」とも言える「出題基準」語彙リストと比較した場合にも明らかで，本研究作成の語彙表は現代日本語書き言葉を理解するための日本語教育語彙表として妥当な語彙選定が行われており，十分に機能することが分かった。さらに，本研究の語彙表の目的は現代日本語書き言葉を理解するために有用な語彙を選

定することであったが，話し言葉テキストにおけるカバー率も書き言葉テキストと同様に高かった。このことから，本研究で作成した語彙表は書き言葉と話し言葉のどちらにおいても有用な語彙を選定することができたと考えられる。

第6章　本研究の結論と展望

第 6 章では本研究の結論と今後の課題および展望について述べる。6.1.では本研究の結果の概要をまとめ，考察する。6.2.では今後の課題と展望について述べる。

6.1節　結果の概要と考察

本研究では，日本語教育語彙表をコーパスと統計指標に基づき客観的に語彙選定し日本語教育的観点からの補正を加えるという方法で作成することを目的とし，語彙表作成を行った。これは，特定のテキストジャンルにおける小規模な語彙調査や専門家の主観的選定に基づいて行う従来型の日本語教育語彙表とは，その手法において根本的に異なるものである。また，コーパス準拠の日本語語彙表である松下（2011）や李・砂川（2012）とも異なる。本研究の語彙表は，コーパスの検討，統計指標の利用方法，日本語教育的観点からのリストの補正方法において異なり，既存の語彙表にはない特長を備えるものとなった。以下，各章の概要と結果を簡潔にまとめる。

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，コーパスに基づき統計指標によっ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，それが，既存の語彙表とは異なるものであることについて述べた。

第 2 章では，先行研究を概観した。先行研究では，小規模語彙調査に基づく従来型の日本語教育語彙表について，次に，コーパスに基づく語彙表について，最後に，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問題と，漢字表記の問題について示した。

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，従来型の日本語語彙表は，このような語彙調査をもとに語彙を選定し，専門家の判定方式で主観的に選ばれ，レベル分けなどがされ

たものであった。その中でも，近年まで日本語教育語彙表の「定番」として最もよく使われてきたのが「出題基準」であった。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た。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示された。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であり，専門家判定方式にも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。また，コーパスにおける語彙の安定度や重要度を示す分布統計や有用度についても，研究が進められてきた。そのような英語教育における教育語彙表作成の知見に基づき，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。

しかし，日本語には分かち書きの習慣がなく，どこまでを一語とみなすかという語の単位に関する問題や表記の問題など，独自の複雑さがある。これらについては，過去の語彙調査研究のころから蓄積されてきた研究成果が現在のコーパス研究にも生かされている。

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，コーパスの選定と再構築の方法について述べた。コーパスが日本語教育語彙表の元データとなるものとして適切かどうかを十分に検討し，調整を加えている点は本研究の語彙表の特長とも言える。そして，再構築したコーパスを頻度集計し，分析用基礎統計を算出する方法と，語彙の選定およびレベル分けの方法について記した。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によるサブコーパスから成る。投野・本田（2016）は，BCCWJ 2009 年度版領域内公開データを使用して，媒体ごとの語彙頻度の相関を分析したが，どのテキストにも安定して出現する語が多いのは頻度上位 100 位までであった。本研究でも，約 1000 万語を実験的にサンプリングし，媒体間の語彙の重なりについて調査したところ，全媒体に共通して出現する語彙は，頻度上位 1000 語までで 110 語，頻度上位 10000 語までで 1944 語であった。このことから，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[29]」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 媒体の語数を削減してサブ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，先行研究による既存の語彙表の作成に利用されたものと比較しても最大のコーパスであり，規模としては十分であると考えられる。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。

この再構築したコーパスは頻度集計し，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 DP（Gries, 2008）を使用した。有用度は DP の逆数に頻度の対数を掛けて算出した。DP を逆数にしたのは，DP が 0 以上 1 以下で示され，0 に近いほど分布が安定していることを示す指標であるためである。さらに，ここに単語親密度を付与した。単語親密度は文字音声親密度を利用した。

次に，分析用基礎統計を付与したデータから語彙を選定，ランキングし，レベル分

[29] 白書（OW），国会会議録（OM），ベストセラー（OB），Yahoo!知恵袋（OC），Yahoo!ブログ（OY），検定教科書（OT）

けを行った。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，10000 語レベルの 5 レベル，総語数 1 万語を選定した。語彙は原則的に有用度順にランキングし，単語親密度でフィルター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った。具体的には，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は，単語親密度に関わらず 2000 語レベルに残した。

第 6 章では，このようにして作成した語彙表について評価を行うため，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，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%以上であり，1 万語を選定した日本語教育語彙表としては十分に機能すると評価できる結果となった。

6.2節 本研究の意義および今後の課題

本研究では，大規模コーパスに基づき，語彙の重要度を統計指標によって定量化し，客観的に語彙を選定したうえで日本語教育的観点からの難易度を加味したランク補正を行う方法で，日本語の書き言葉を理解するために有用な 1 万語規模の日本語教育語彙表を作成した。その結果，選定した語彙のテキストカバー率は高く，日本語を理解するために適切な基本語彙を選定することができた。本研究で作成した語彙表は様々

な利用可能性がある。日本語教育語彙表として新しい試みも行った。しかし，いくつかの課題も残されている。

まず，利用可能性としては，以下の(1)~(4)のような教材やテストへの利用が考えられる。

(1) 単語集や語彙問題集などの形で，日本語を読んで理解するための重要語彙を学習することを目的とした教材に加工。
(2) 読解教材のリライトやグレーディッドリーダーズの作成などで語彙を制限するための資料として利用。
(3) 読解テキストの語彙的な難易度を測定する基準として利用。
(4) 書き言葉の日本語を読んでどのくらい理解できるかを語彙の側面から測定するテストの出題語彙として利用。

本研究の語彙表は，統計指標に基づく重要度だけでなく，日本語教育的視点からの難易度も加味したランク調整を行っている。その結果，上記のような利用方法に適したものになっている。

(1)と(2)は教材利用である。日本語教育では，コーパス準拠で選定した語彙を加工して教材にしている例はまだ多くない（深田，2008，中條，2009）。例えば，(1)に関しては，近年出版されたいわゆる単語集で一般的に利用されているものに『日本語能力試験ターゲット 2000（改訂版）』シリーズ（旺文社，2015）や『日本語単語スピードマスター』シリーズ（J リサーチ出版，2010-2015）などがある。『日本語能力試験ターゲット 2000（改訂版）』はタイトルの通り試験対策教材で日本語能力試験過去五年分の試験問題を分析して語彙を選定している。『日本語単語スピードマスター』は試験対策だけに焦点を当てたものではないが，収録されている語彙は「日本語能力試験の出題基準など，さまざまな資料を参考に，生活でどのように使われているかを考えて」選定したと説明されている。これらは一例ではあるが，日本語教育では単語集で

も語彙問題集でも日本語能力試験を視野に入れ，日本語能力試験の過去問題や「出題基準」を語彙選定の主な基準としているものが多い。日本語能力試験は日本語学習者にとって重要な学習動機の一つであり，これに準拠した教材が数多く出版されているのは自然なことである。しかし，日本語能力試験以外のデータ，すなわちコーパス準拠で客観的に語彙を選定し一般的な日本語のテキストを読むために重要な語彙を集めた教材も必要である。また，語彙知識には受容語彙と産出語彙，話し言葉と書き言葉など複数の側面があるが，これらのどのような側面に効いてくるのかという視点から編集された教材があってもよい。語彙のテキストカバー率調査の結果，本研究で選定した語彙は話し言葉コーパスにおけるカバー率も高かったので，本研究で選定した語彙を学習することは日本語の書き言葉だけでなく話し言葉の理解にも役立つ。(2)に関しては，NPO 多言語多読という団体によって日本語教育多読教材が作成されている。多読教材は独自に開発された語彙表に基づきレベル分けがされており，この語彙表も「出題基準」をベースに作成者の「長年の経験」から語彙選定を行ったと記されている。語彙の難易度はリーダビリティに関わる重要な要素であるが，多読教材の語彙レベルも日本語能力試験と選定者の主観だけで決められているのが現状である。したがって，本研究で選定したような生の日本語テキストの理解において必要な語彙とそのレベル分けは，ここでも利用価値が高いものと考えられる。これは読解教材のリライトにおいても同様である。

(3)と(4)は主にテスティングへの応用である。(3)はリーダビリティに関わる。本研究の語彙表はテストや読解教材を作成する際，読解テキストで使用されている語彙の難易度を推定する資料としても有用である。特に，本研究の語彙表は，日本語の書き言葉を理解する際に必要な語彙を選定しレベル分けを行っているので，このような目的で作られていない他の既存の語彙表などと比べても，読解テキストの語彙の難易度判定に適している。(4)は語彙知識を問うテストでも特に受容語彙の知識を測定するものである。本研究の語彙表は，学習者がどのような順番で語彙を習得するのかを調査したものではないので，学習者が読んで理解できる語彙が何語程度なのかを測定する

ような，いわゆる語彙サイズテストには使うことはできない。しかし，学習者が日本語の書き言葉の語彙をどのぐらい理解できるのか，また，語彙知識に基づきどの程度のテキストが読めるのか，ということは測定可能である。ただし，日本語の書き言葉を理解するためには表記，すなわち漢字の難易度も大きな要素となるが，ここでは語彙の意味の側面にのみ焦点を当てている。

次に，本研究で作成した語彙表の日本語教育研究における意義について述べる。まず,本研究ではコーパスを利用するにあたりコーパス自体の評価や検討を十分に行い，最適化した形で使用した。これまでの日本語教育語彙表ではコーパスを使っていてもコーパスの中身から検討した例はない。コーパスは基本語彙の選定において利用価値の高いものであるが，コーパスにおける語彙頻度情報はそれを構成するテキストジャンルの影響をそのまま受けるものである。したがって，どのようなテキストジャンルのサブコーパスで構成されているか，作成する語彙表の目的に合っているかという点を検討することは非常に重要である。

そして，語彙の選定は統計指標や単語親密度などを利用して客観的に，かつ日本語教育的な難易度も付加する形で行った。日本語教育の語彙表では，専門家判定方式か，「出題基準」のような専門家判定方式で作られた語彙表を参考に語彙のランキングやレベル分けが行われる例がほとんどで，それだけ専門家判定方式の信頼が厚いとも言える。しかし，そのような語彙表で選定された語彙表間の一致率が高くないという問題（2.1.4.参照）を受け，本研究では可能な限り客観的に語彙を選定するよう努めた。

さらに，本研究の語彙表は「一般成人日本語学習者に向け，書き言葉日本語を理解するために有用な語彙を選定する」とし，その目的を明示している。従来の語彙表は対象者まではっきりしていても,その語彙が有用であるのは話し言葉か書き言葉か，それらは学習すべき語彙か語彙習得面を考慮しているのかなどの点がなどを明示していないものも少なくない。本研究では，日本語の書き言葉コーパスを利用して，特に書き言葉を理解するための語彙を抽出した。結果として，話し言葉にも対応できる語彙表となったが,このように語彙表の利用可能性を絞って基本語彙を提示することは，

学習面でも研究においても有意義である。

しかし，今後の課題として，複合語の問題，表記の問題，日本語教育的観点からの補正に関する問題，中頻度以降の語彙レベルの問題，他のコーパス準拠の語彙表との一致率調査が残った。

まず，本研究では，語の単位を短単位で処理し，語彙素の表記のまま見出し語項目とした。これには次のような理由がある。まず，短単位はレマであり語彙学習における認知的な負荷を表すのに適している（2.2.2.参照）。例えば，短単位の「積極」という項目は「積極的」や「積極性」などの派生語を含むことになる。「～的」「～性」という接尾語の知識があれば，認知的な負荷は「積極」の部分だけのものになる。そのような意味で，本研究で作成する語彙表の目的である「現代日本語書き言葉を読むための語彙」の知識を示すものとして，短単位は適している。しかし，日本語教育的には通常「積極」だけの使用はせず「積極的」「積極性」といった派生語の形で使うものであるという記述が，特に学習者向けに加工する場合においては必要である。また，合成語も高頻度のものや一般的なものは一定量取り出せるのが望ましい。語彙表は重要な複合語やコロケーション情報を提示することにより，語彙チャンクやコロケーションフレーズを教え込むレキシカルアプローチ（Lexcal Approach）へと利用範囲が広がる。なお，このような複合語の問題について，李・砂川（2012）では，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ている。また，Tono et al. (2013)では，長単位を採用し，複合語を含めた語彙の頻度と分布を体系的に示している。今後はこのような研究を参考に，日本語教育的観点から重要と考えられる複合語やコロケーションの抽出も行いたい。

次に，表記の問題であるが，本研究の語彙表の見出し語表記は，Unidicの語彙素表記をそのまま使用した。語彙素表記は，同音異義語が多い日本語の性質上，通常の書き言葉では使われないような漢字表記が多く見られる。敢えてこの形のままとしたのは，二つの理由がある。一つは，今後，本研究の語彙表を研究用などに利用する場合，語彙素表記のままのほうが使いやすいと考えたためである。例えば，テキストカバー

率調査をする場合に，テキストを形態素解析して行えば分析が容易である。日本語の表記の問題は複雑で，同じ語であっても，テキストによって漢字と平仮名で違っていたり，同じ漢字でも送り仮名の付け方が違っていたりする。二つ目は，どのような表記を日本語教育的な観点で「標準的」とするかが難しいからである。さらに，日本語教育的には，漢字の難易度もまた別の問題として存在する。しかし，日本語教育の現場や学習者の利用を考えた場合，表記辞典などを参考に標準的な表記も示したほうがよい。この点は今後の課題として残った。

また，本研究では，日本語教育的観点から見た語彙のランキングの順位補正を高頻度語彙に限定して必要最低限の範囲で行ったが，今後はより詳細な調査に基づき，語彙のランキングやレベル調整をしていきたい。本研究ではコーパスの出現頻度と分布統計に基づき語彙の重要度を定量化したうえで，日本語教育的な難易度を判定するために単語親密度でそれらの語彙をフィルターした。そして，高頻度語彙に限定して，初級レベルの学習において必ずといってよいほど出現する語彙が単語親密度「クセ」の影響によって下位のレベルに配置されていればそれを元の位置に繰り上げるという順位補正を行った。しかし，もっと詳しく複数の日本語教科書で扱われる文型とそれに結びつきやすい語彙を体系的に調査していけば，順位調整が必要な語彙がさらに出てくるかもしれない。今後はこのような語彙に関してさらに詳しく調査し，必要に応じて調整をしていきたい。

さらに，今回は総語数 1 万語規模で語彙表を作成したが，この規模に関しては十分な根拠があるとは言えない。語彙の出現頻度はコーパスの影響を受けるものであり，頻度が低いものほどコーパスの違いで頻度ランクに大きな違いが出てくる。特に，頻度上位 5000 語以降あたりからはその違いが顕著である(Tono, 2013)。今後は，特に中上級レベルの語彙の検証を複数のコーパスから選定した語彙との比較において行っていきたい。

最後に，既存のコーパス準拠の語彙表との比較，語彙の一致率調査を課題として挙げたい。本研究の語彙表の目的と類似する既存のコーパス準拠の語彙表の主なものに

は松下（2011）と李・砂川（2012）があるが，使用したコーパス，語彙の選定，レベル分け方法などが異なっている。今後は，本研究の語彙表とこれら二つの日本語教育語彙表の語彙の一致率はどの程度か，類似点，相違点，それぞれの特長は何かなどの点について調査，分析を行いたい。

以上が本研究の今後の課題として残った。今後はこれらの問題に取り組み，本研究の語彙表の改善に努め，日本語教育語彙表の研究をさらに進めていきたい。

謝辞

本研究を進めるにあたり，ご指導をいただいた指導教員の東京外国語大学総合国際学研究院投野由紀夫教授に感謝を申し上げます。懇切丁寧に指導してくださったことはもちろんのこと，日常の様々な問題に直面したときにもサポートしてくださいました。特に，博士後期課程に進んでからは結婚，海外生活，健康上の問題，出産などで研究だけに集中できない時期もありましたが，そのような生活上の問題にも理解を示してくださり，寛容に論文の原稿が書き上がるのを待ってくださいました。投野先生の下で学び，博士論文を執筆できたことは大変幸せでした。

また，貴重なコメントをくださった東京外国語大学総合国際学研究院佐野洋教授にも感謝いたします。本研究の意義や分析の方法について，異なる視点から貴重なご意見をいただきました。また，Word の使い方といった基本的で細かい点まで，丁寧にご指導いただきました。

そして，日常の議論を通じて多くの知識や示唆をいただいた投野ゼミの皆様に感謝します。特に，金田拓氏と村上明氏には深く御礼を申し上げます。

最後に，いつも温かく見守り，協力してくれた家族に感謝します。産まれたばかりの息子の存在は，博士論文を仕上げる気力の源となりました。

本研究を支えてくださったすべての方々に深い感謝を捧げます。

参考文献

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing 8 (4)*, 243 – 257.

Caroll, B. J. (1970). An Alternative to Juilland's Usage Coefficient for Lexcal Frequencies and a Proposal for a Standard Frequency Index (SFI). In *Computer studies in the Humanities and Verbal Behavior, 3(2)* (pp. 61-65).

Caroll, J. B., Davies, P., & Richman, B. (1971). *The American Heritage Word Frequency Book.* Boston: Houghton Mifflin Co.

Francis, H., & Kučera, N. W. (1982). *Frequency Analysis of English Usage.* Boston: Houghton Mifflin Company.

Francis, N. (1982). Problem of assembling and computerizing large corpora. In I. S. (Ed.), *Computer corpora in English language research* (pp. 7-24). Bergen, Norway: Norwegian Computing Center for the Humanities.

Gries, S. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics 13 (4)*, 403 – 437.

Hofland , K., & Johasson, S. (1982). *Word frequency in British and American English.* Bergen, Norway: The Norwegian Computing Center for the Humanities.

Juilland, A., Brodin, D., & Davidovitch, C. (1970). *Frequency Dictionary of French Words.* Hague: Mouton.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence: RI: Brown University Press.

Leech, G. (1992). Corpus and theories of linguistic performance. In J. Svartik, *Directions in corpus linguistics* (pp. 105-122). Berlin, Germany: Mouton de Gruyter.

LeechG., RaysonP., WilsonA. (2001). Word frequency in written and spoken English: Based on the British National Corpus. . Harlow, UK: Pearson Education Limited.

Nation, I. (2001). *Learning vocabulary in another language.* Cambridge, UK: Cambridge University Press.

Nation, I.S.P. (2008). *Teaching Vocabulary Strategies and Techniques.* Boston, USA: Heinle.

Neustupný, J. V. (1977). *A Classified List of Basic Japanese Vocabulary.* Melbourne: Monash University, Department on Japanese.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford, UK: Oxford University Press.

Swenson, E., & West, M. P. (1934). On the counting of new words in textbooks for teching foreign languages. In *Bulletin of the Department of Educational Research.* University of Tronto, 1.

Thorndike, L. E., & Lorge, I. (1944). *The teachers word book of 30,000 words.* New York: Teachers College. Columbia University.

Tono, Y. (2013). Sampling biases and implications for better wordlist creation. *Vocab@Vic conference, presentation slides.* Victoria University of Wellington.

Tono, Y., Maekawa, K., & Yamazaki, M. (2013). *A Frequency Dictionary of Japanese.* London: Routledge.

West, M. (1953). *A General Service List of English Words.* London: Longman, Green and Co.

Zeno, S., Ivens, S., Millard, R., & Duvvuri, R. (1995). *The Educator's Word Frequency Guide.* Brewster, NY: Touchstone Applied Science Associates.

Zipf, G. K. (1935). *The Psycho-Biology of Language.* Boston-Cambridge Mass:

Houghton Mifflin.

Zipf, G. K. (1949). *Human Behavior & The Principle of Least Effort, An Introduction to Human Ecology.* Addison-Wesley Press Inc.

饗場淳子.(2011).「日本語教育用語彙に共通する語についての一考察」. 早稲田大学大学院教育学研究科紀要 18-2, 275-285.

秋元美晴・押尾和美.(2008).「新しい日本語能力試験のための語彙表・漢字表作成中間報告--新語彙表 ver.3 の完成まで 」.『日本語学』27(10), 36-49.

秋元美晴.(2002).『よくわかる語彙』. アルク.

天野成昭, 近藤公久.(1999).『NTT データベースシリーズ　日本語の語彙特性（第 1 期)』. 三省堂.

池原楢雄.(1957).『国語教育のための基礎語彙体系』. 六月社.

石川慎一郎, 前田忠彦, 山崎誠, （編）.(2010).『言語研究のための統計入門』. くろしお出版.

石川慎一郎.(2008).『英語コーパスと言語教育』. 大修館書店.

石川慎一郎.(2009).「日本語基本語研究における非統制型・統制型・媒介型 Web as Corpus の可能性 ―言語コーパスにおける基本語頻度の安定性について」. 特定領域研究「日本語コーパス」サテライトセッション予稿集, 29-38.

夷石寿賀子, 庄寿千葉, 陳君慧.(2006).「『青空文庫』を言語コーパスとして使おう―メタデータ構築による歴史的・社会言語学的研究への応用の試み―」. 言語処理学会第 12 回年次大会（NLP2006）発表論文集, 915-918.

宇佐美まゆみ（監修）.(2011).『BTS による多言語話し言葉コーパス 2011 年版』. 東京外国語大学宇佐美まゆみ研究室.

岡本禹一.(1944).『日本語基本語彙』. 国際文化振興会.

小椋秀樹.(2006).「形態論情報」.『日本語話し言葉コーパスの構築法』国立国語研究所報告 124, 133-180.

小椋秀樹.(2007).「『現代日本語書き言葉均衡コーパス』における短単位の概要」. 特

定領域研究「日本語コーパス」平成 18 年度公開ワークショップ（研究成果報告会）予稿集, 101-108.

小椋秀樹. (2009). 「コーパスのための形態論情報」. 『国語学　解釈と鑑賞』74 巻 1 号, 26-34.

押尾和美, 秋元美晴, 武田明子, 阿部祥子, 高梨美穂, 柳沢好昭, . . . 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』第 4 号, 71-86.

甲斐睦朗. (2002). 「現代日本語の基本語彙」. 著: （編）飛田良文・佐藤武義, 『現代日本語講座 第 4 巻 語彙』 (ページ: 25-45). 明治書院.

加藤彰彦. (1963-34). 「日本語教育における基礎学習語」. 『日本語教育』2, 4, 5 号.

樺島忠夫・吉田弥寿夫. (1971). 「留学生教育のための基本語彙表」. 著: 『日本語・日本文化』2. 大阪外国語大学留学生別科.

河原大輔・黒橋禎夫. (2006). 「高性能計算環境を用いた Web からの大規模格フレーム構築」. 情報処理学会研究報告自然言語処理（NL）, 67-73.

川村よし子. (2006). 「日本語学習者のための基本語選定の一試案」. 『ヨーロッパ日本語教育』vol.11, 72-78.

北原保雄. (2002). 『朝倉日本語講座 4　語彙・意味』. 朝倉書店.

北原保雄. (2005). 『朝倉日本語講座 2　文字・書記』. 朝倉書店.

工藤真由美. (1999). 『児童生徒に対する日本語教育のための基本語彙調査』. ひつじ書房.

現代日本語研究会（編）. (2002). 「男性の言葉・職場編」. ひつじ書房.

現代日本語研究会（編）. (1998). 「女性の言葉・職場編」. ひつじ書房.

小磯花絵. (2009). 「話し言葉コーパスの情報」. 『国語学　解釈と鑑賞』74 巻 1 号, 53-60.

国語学会（編）. (1980). 『国語学大辞典』. 東京堂出版.

国際交流基金・日本国際教育協会. (2002). 『日本語能力試験出題基準（改訂版）』. 凡

人社.
国立国語研究所.(1970).『電子計算機による新聞の語彙調査』.秀英出版.
国立国語研究所.(1979).『日本語教育語彙資料(1)(2)―低学年初級 500 語』.国立国語研究所日本語教育第二研究室.
国立国語研究所.(1982).『日本語教育基本語彙七種比較対照表』.大蔵省印刷局.
国立国語研究所.(1982).『日本語教育指導参考書 9 日本語教育基本語彙七種比較対象表』.大蔵省印刷局.
国立国語研究所.(1984).『高校教科書の語彙調査Ⅱ』.秀英出版.
国立国語研究所.(1984).『国立国語研究所報告 78 日本語教育のための基本語彙調査』.秀英出版.
国立国語研究所.(1984).『日本語教育のための基本語彙調査』.秀英出版.
国立国語研究所.(1987).『中学校教科書の語彙調査Ⅱ』.秀英出版.
国立国語研究所.(2003).『語彙の研究と教育(上)』.大蔵省印刷局.
国立国語研究所.(2006).『日本語話し言葉コーパス』.国立国語研究所.
国立国語研究所.(2011).『現代日本語書き言葉均衡コーパス』.国立国語研究所.
国立国語研究所(編).(2004).『分類語彙表 増補改訂版』.大日本図書.
国立国語研究所コーパス開発センター.(2011).『現代日本語書き言葉均衡コーパス』利用の手引き 第 1.0 版.国立国語研究所.
国立国語研究所日本語教育センター第二研究室分室.(1992).「簡約日本語」.著:『簡約日本語の創成と教材開発に関する研究』.国立国語研究所日本語教育センター第二研究室分室.
近藤安月子・小森和子(編).(2012).『研究社日本語教育事典』.研究社.
坂本一郎.(1943).『日本語基本語彙 幼年之部』.明治国書.
坂本一郎.(1984).『新教育基本語彙』.学芸図書.
佐藤政光.(1999).「日本語学習者の語彙習得に関する調査研究―(1)基本語彙の問題点について」.『明治大学人文科学研究所紀要』第 44 冊, 169-180.

志部昭平（編）.(1980).『日本人の知識階層における話し言葉の実態』. 国立国語研究所.ジョウホウキョク情報局.(1942).『簡易基本ニッポン語』. 日本読売新聞.
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一『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.
専門教育出版『日本語学力テスト』運営委員会.(1998).『品詞別・A~D レベル別 1 万語語彙分類集』. 専門教育出版.
田中久直.(1956).『学習基本語彙』. 新光閣書店.
玉村文郎.(1970, 1978). Practical Japanese－English Dictionary. 海外技術者研修協会.
玉村文郎.(1987).「日本語教育基本 2570 語」. 著:『NAFL Institute 日本語教師養成通信講座 8 日本語の語彙・意味』. アルク.
玉村文郎.(2003).「中級用語彙--基本 4000 語」.『日本語教育』116 号, 5-28.
玉村文郎（編）.(1989).『講座日本語と日本語教育 第 6 巻日本語の語彙・意味（上）』. 明治書院.
中條清美.(2009).「コーパスを活用した日本語教材作成の試み」. 日本大学生産工学部研究報告 B 第 42 巻, 43-52.
中條清美.(2009).「指導に役立つ語彙リスト紹介」.『G.C.D.英語通信』No.46, 10-11.
土居光知.(1933).『基礎日本語』. 六星館.
投野由紀夫, 本田ゆかり.(2010).「基本語彙の頻度と分布統計を用いた BCCWJ のサンプリングの評価と分析：中間報告」. 2010 年 5 月 8 日日本語教育班会議資料.
投野由紀夫, 本田ゆかり.(2016).「第 2 章 教育語彙表への応用」. 著: 砂川有里子(編),『講座日本語コーパス 5 コーパスと日本語教育』. 朝倉書店.
投野由紀夫.(2009).「コーパスと英語教育」.『国語学 解釈と鑑賞』74 巻 1 号, 95-103.
徳永健伸（著）, 辻井潤一（編）.(1999).『言語と計算 5 情報検索と言語処理』. 東京大学出版.

徳弘康代.(2005).「中上級学習者のための漢字語彙の選択とその提示法の研究」.『日本語教育』127 号, 41-50.

日本語教育学会（編）.(1991).「初級日本語教科書によく使われる語」. 著: 日本語教育学会（編）,『日本語教育機関におけるコースデザイン』. 凡人社.

橋本直幸・山内博之.(2008).「日本語教育のための語彙リストの作成」.『日本語学』27(10), 50-57.

橋本直幸.(2010).「日本語教育語彙リストのための話題特徴語の抽出」.『特定領域研究「日本語コーパス」平成 22 年度全体会議予稿集』, 296-301.

深田淳. (2008).「コーパス言語学の日本語研究・日本語教育への応用」. Fifteenth Princeton Japanese Pedagogy Forum PROCEEDINGS（ページ: 1-18). http://www.princeton.edu/pjpf/2008proPDF/5%20Fukada_PJPF08.pdf. 参照先 : Fifteenth Princeton Japanese Pedagogy Forum PROCEEDINGS,http://www.princeton.edu/pjpf/2008proPDF/5%20Fukada_PJPF08.pdf.

文化庁国語課.(1971, 1975).『外国人のための基本語用例辞典』. 大蔵省印刷局.

本田ゆかり.(2009).「大規模コーパスを用いた日本語学習語彙表作成の試み」.『特定領域研究「日本語コーパス」平成 21 年度全体会議予稿集』, 127-136.

前川喜久雄, 山崎誠.(2009).「『現代日本語書き言葉均衡コーパス』.『国語学　解釈と鑑賞』74 巻 1 号, 15-25.

前川喜久雄.(2009).「コーパスとは何か」.『国文学　解釈と鑑賞』, 6-14.

松下達彦.(2010).「日本語を読むために必要な語彙とは？－書籍とインターネットの大規模コーパスに基づく語彙リストの作成ー」.『2010 年度日本語教育学会春季大会予稿集』, 335-336.

松下達彦.(2011).「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.

丸山岳彦. (2009).「日本語コーパスの現状」.『国語学　解釈と鑑賞』74 巻 1 号,

122-130.

望月正道・相沢一美・投野由紀夫.(2003).『英語語彙の指導マニュアル』.大修館書店.

山内博之.(日付不明).「KY コーパス」.参照先: 日本語 OPI 研究会ウェブサイト「OPI を利用したコーパス」: http://www.opi.jp/shiryo/ky_corp.html

山内博之（編）.(2008).『日本語教育スタンダード試案 語彙』.ひつじ書房.

山崎誠.(2006).「国立国語研究所の語彙調査の歴史と課題」.東京大学大学院教育学研究科教育測定・カリキュラム開発（ベネッセコーペレーション）講座, 第 12 回公開研究会資料.

山崎誠.(2009).「国立国語研究所における諸研究」.『国語学　解釈と鑑賞』74 巻 1 号, 183-191.

山下喜代, 秋元美晴, 小宮千鶴子.(2008).『日本語教育のための合成語データベース構築とその分析』.平成 17 年度~19 年度科学研究費補助金基盤研究（C）研究成果報告書.

李在鎬・石川慎一郎・砂川有里子.(2012).『日本語教育のためのコーパス調査入門』.くろしお出版.

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ー現状と今後の展望ー（名古屋大学）.

李在鎬.(2013).「大規模コーパスに基づく語彙リストの検証」.日本語学習辞書科研マレーシア研究集会資料.

Sketch Engine. (n.d). Retrieved from http://www.sketchengine.co.uk/

青空文庫.(日付不明). 参照先: http://www.aozora.gr.jp/

アルク.(日付不明). アルク「標準語彙水準 SVL12000」(SVL＝Standard Vocabulary List). 参照先: http://www.alc.co.jp/eng/vocab/svl/

内山将夫, 高橋真弓. (2003). 「日英対訳文対応付けデータ」. 参照先: http://www2.nict.go.jp/univ-com/multi_trans/member/mutiyama/align/index.html

カドカワ角川書店. (日付不明). Web KADOKAWA. 参照先: http://www.kadokawa.co.jp/book/tachiyomi.html

上村隆一. (1998). 「インタビュー形式による日本語会話データベース（上村コーパス)」. 参照先: http://www.env.kitakyu-u.ac.jp/corpus/

幻冬舎. (日付不明). 幻冬舎 Plus. 参照先: http://www.gentosha.jp/

国際交流基金・日本国際教育支援協会. (2012). 「日本語能力試験 JLPT」. 参照先: 「旧試験との比較」: http://www.jlpt.jp/about/comparison.html

国立国会図書館. (日付不明). 国会会議録検索システム. 参照先: http://kokkai.ndl.go.jp/

ポプラ社. (日付不明). ポプラビーチ. 参照先: http://www.poplarbeech.com/

松下達彦. (2011 年 12 月 18 日). 「日本語を読むための語彙データベース」(The Vocabulary Database for Reading Japanese). 参照先: http://www17408ui.sakura.ne.jp/tatsum/LTVJ/index.html