日本語教育用の形容詞の語彙リストと難易度レベル

スルダノヴィッチ・イレーナ（国立国語研究所日本語教育研究・情報センター／
リュブリャーナ大学文学部）†
李在鎬（筑波大学人文社会系）

Vocabulary List of Adjectives and Levels of Difficulty for Japanese Language Education

Irena Srdanović (National Institute for Japanese Language and Linguistics/
University of Ljubljana)
Lee Jae-Ho (University of Tsukuba)

１．はじめに

大規模コーパスの構築と共に、コーパスに現れる語彙の把握ができるようになり、バランスが取れたコーパスほど、抽出された高・中・低頻度の語彙が実際に利用される語彙の実情を示す傾向が見られる。どのコーパスでもそれぞれの独自の特徴があり、その特徴は語彙分布にもあるが、コーパスが大ければ大きい、また均衡が取れていればそれだけ、語彙の分布に偏りが少なく、得られた語彙データの信頼性が高くなる。大規模コーパスにおいてサブコーパス別、いわゆるジャンル別のデータも得られるようになり、分散度（dispersity）による語彙の特徴が取り出せるようになってきた。第二言語教育においてもこのような語彙リストがよく使われるようになり、近年複数のリソースを利用して作成されている。[1]Nation（2001）によると、英語の高頻度の 2000 基本語彙がテキストの内容 70%～80%をカバーするため、学習者にまずその語彙を教えるべきという指摘がある。

従来の国語研究および日本語研究において、語彙リストの研究は様々なものあり、語彙を確定するために、語彙頻度調査の実施、専門家の判定、編者の判定、児童・生徒の理解度の調査、成人の獲得語数の調査、語の親密度の調査などの方法が用いられてきた。また、『教育基本語彙の基本的研究』のような複数の語彙リストがすでにデータベース化されている。そのうち日本語学習者を対象にしたリストの例としては、「日本語教育基本語彙データベース」（教育基本語彙の基本的研究―増補改訂版 2008）、「日本語能力試験出題基準」（国際交流基金・日本国際教育協会 1994）があげられる。大規模な現代日本語書き言葉均衡コーパス（以下 BCCWJ）などのコーパス開発と共に、コーパスを基にした日本語教育向けの語彙リストの作成が始まった。その例は、近年作成された「日本語を勉強する人のための語彙データベース」（松下 2011）および「日本語教育語彙表」（砂川 2012、李・砂川 2012）である。

日本語教育用の語彙リストはいくつかあるが、その作成方法、基にした資料の特徴・現代性などに違いがあり、どの程度収録された語彙が一致しているかについては必ずしも明らかではない。饗場（2011）が 3 種の語彙リストを調べた結果、リストごとに非共通語が多くあると明らかにした。例えば、形容詞・形容動詞を取り上げると、各語彙における共通語彙の割合は 54.8%、59.2%、90.5%である。スルダノヴィッチ（2012）がリストごと、コーパスごとに形容詞の語数を比較した結果、差異が多く見られることが分かった。本論文の目的は、形容詞を対象にした既存の日本語の語彙リストを検討し、その語彙リストにあ

† irena.srdanovic@ff.uni-lj.si

1　例えば、日本人英語学習者のための語彙リスト JACET8000（大学英語教育学会基本語改定委員会(編) 2003）は、学習者が遭遇しやすいサブコーパスと BNC コーパス頻度を対数尤度比で比較し、作成されたものである。

る形容詞と大規模コーパスから取り出せる形容詞を比較しつつ、語彙リストに把握されていない項目を検討することにある。そこから得られたデータを基にして、今後の課題としては、新たな日本語学習者用形容詞の語彙リストおよび形容詞と他の単位との組み合わせの記述を目指すことである。

２．日本語教育用の語彙リスト

以下に取り上げる語彙リストは、本研究の対象にして、それぞれのリストに現れる形容詞を検討する。

２．１『日本語能力試験出題基準』の旧語彙リスト

『日本語能力試験出題基準』（1994）の旧語彙リスト（以下「旧 JLPT 語彙リスト」）はテスト作成のために作られたものであり、教育目標のために作成されたリストではないが、日本語教育において幅広く利用されている。語彙難易度は 4 段階に分かれ、下位級の 4 級から上位級の 1 級までである。旧 JLPT 語彙リストは、作られてから 30 年以上経過しているため、語彙の変化に対応していない、カタカナ語や擬音語・擬態語などの語彙が少ないという問題点がある（李・砂川 2012）。なお，2010 年から実施されるようになった新しい日本語能力試験のために作られた「語彙リスト」は 5 段階の難易度に分けられているが，テスト運用上の理由から非公開になっている。

２．２　「日本語教育基本語彙データベース」、「教育基本語彙データベース」

「日本語教育基本語彙データベース」（以下「国研日本語語彙 DB」）は、「国語研教育基本語彙データベース」に登録した 6103 語、国立国語研究所『日本語教育基本語彙七種比較対照表』の 6195 語などの 6 種の教育基本語彙リストのデータをデータベース化したものである。データベース化された 6 種のリストは様々な方法で集められた語彙データであり、総数は 11826 語である。その詳細は国立国語研究所報告 127『教育基本語彙の基本的研究』（2009、 563 ページ）において確認することができる。

同じく国立国語研究所報告 127 に搭載されている「教育基本語彙データベース」（以下「国研国語語彙 DB」）は、7 種の教育語彙を利用したデータベースで、主に国語教育のデータをカバーしている。語数は、27234 である。データベースは小学生低学年・高学年、中学生の理解度を測定したデータに基づいて語彙難易度の 3 段階に分けている（1 は最も低い難易度）。

２．３　「日本語を勉強する人のための語彙データベース」

「日本語を勉強する人のための語彙データベース」（以下「TM 語彙リスト」）は、『現代日本語書き言葉均衡コーパス』（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）を使って、松下（2010、2011）が作成した語彙リストであり、Nation（2001）の英語学習のために提案された語彙リスト作成の枠組みに基づいている。その特徴は、コーパス頻度およびサブコーパスごとの語彙分布を基にして、語彙を「一般用」、「留学生用」に分けたデータである。一般用のデータは基本 2500 語を含み、総合数は 20326 語である。留学生用のデータは 3・4 種の分野でよく使われる単語であり、科学分野別の特徴のあるデータも掲載されている（20312 語）。語彙レベルで一般人の生活を中心に考えた重要度のランクおよび Basic、Inter、Adv、H-Adv、S-Adv 五つの語彙ランクがある。語彙の見出し語には UniDic 辞書の短単位の「語彙素」を使っている。リストには旧 JLPT 語彙リストの語彙難易度、語種、品詞などの情報が含まれている。

２．４　「日本語教育語彙表」

「日本語教育語彙表」は、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 18010 語である。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の2009年度版の公開データを利用し、見出し語を決定している。また見出し語に対して，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し、統計的に調整したデータである。見出し語にはUniDicに基づく短単位と単語N-gramによる複合語が入っている。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。

２．５　その他

上述した日本語教育用の語リスト以外に、話題別語彙表などがある（山内（編）2008、橋本・山内2008）。また、近年Can-doタスクに基づいた語彙表の作成が行われている。現在の段階でタスク・話題に関するデータは少ない。課題遂行能力に基づくコミュニケーションのための日本語教育のためには、山内(編)(2008)のような試みは今後も加速化されるべきであろう。

３．コーパスから取り出せる語彙リスト

上に取り上げた日本語教育語彙表とTM語彙リストは、コーパスから取り出した語彙に基づいて作成されたデータであるが、両方のリストはBCCWJの全体版が公開されていないときに作られたものである。本研究では、BCCWJの全体コーパスおよび大規模なウェブコーパスから取り出した語彙頻度リストを利用し、既存のデータと比較する。

BCCWJは総語数1億語の大規模コーパスで、次のサブコーパスで構成されている：出版書籍（PB）、出版新聞（PN）、出版雑誌（PM）、図書館書籍（LB）、また特定目的コーパスとして白書（OW）、ベストセラー（OB）、知恵袋（OC）、ブログ（OY）、法律（OL）、国会会議録（OM）、広報誌（OP）、教科書（OT）、韻文（OV）である。本研究ではMeCabとUniDicの短単位で解析されたデータを利用した。

JpWaCは4億語のウェブコーパスで、スケッチエンジンというレクシカルプロファイリングツールに載せている（スルダノヴィッチ・仁科 2008）。このコーパスは、副詞分布による13種のデータを分析した結果、均衡BCCWJコーパスの書籍のデータに最も類似しており、偏りの少ないデータであることが明らかになった（Srdanovićら2008）。コーパスはChaSen と IPADICで解析されたデータで、統一するため、取り出した形容詞のリストをUniDicで再解析する。

４．日本語教育用の形容詞の語彙リスト

本節では既存リストおよびコーパスに現れる形容詞の項目を調べ、新しい日本語教育用の形容詞語彙リストの項目として検討する。クラスター分析でコーパスごとに現れている最も高頻度の形容詞を分析し、グラフで表示した上でその形容詞の日本語教育においての利便性を議論する。

４．１ 既存の語彙リストに現れる形容詞—語数と難易度

表1は既存の語彙リストに現れる形容詞の語数を示している。

コーパス頻度と分散度に基づいたTM語彙リスト（一般基本語彙）にある形容詞は93語である。この形容詞は最も基本的な語彙で、高頻度であり、様々なサブコーパスに表れるので早い段階で導入し、学習するようにリストの作成者が推薦している。国研国語語彙DBは日本人の国語教育用のため、形容詞数は大きくなっている（460語）。他のリストは日本語教育用の語彙リストで、いわゆる日本語学習者が勉強するための語彙がカバーされており、高い数字からみると、TM一般語彙リスト（353語）、TM留学生語彙リスト（345語）、日本語語彙表（302語）、国研日本語語彙DB（236語）、旧JLPT語彙リスト（245語）で並ぶ。このうちもっとも収録語数が多いリストと収録語数が少ないリストを見ると3分の1の差が見られる。

表 1　語彙リストに現れる形容詞の語数

	旧 JLPT 語彙リスト[1]	国研 国語 語彙 DB	国研 日本語 語彙 DB	TM 語彙リスト（一般）	TM 語彙リスト 2500	TM 語彙リスト（留学）	日本語教育語彙表[1]
形容詞-一般	245	460	236	342	90	334	280
接尾辞-形容詞的	/	/	/	8	0	8	12
形容詞-非自立可能	/	/	/	3	3	3	10
合計 - 形容詞 - 語数	245	460	236	353	93	345	302
合計 - 形容詞 - %	3, 27	1, 69	2, 00	1, 74	3, 68	1, 70	1, 68
合計 - 語彙リスト - 語数	7500	27234	11826	20326	2524	20312	18011

[1] 30 語の形容詞は他の品詞と形容詞と両方分析されている語も含んだ。

更に語彙リストにおける形容詞の難易度レベルを調べた結果、図 1 に示した。

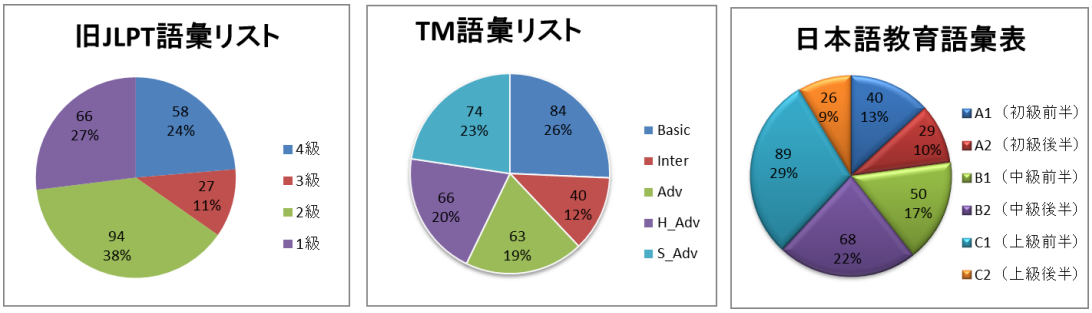

図 1　語彙リストにおける形容詞の難易度―レベルごとの形容詞の語数・割合

旧 JLPT 語彙リストでは 2 級の形容詞は 40%ぐらいの形容詞となっており、その妥当性に疑問が残る。TM 語彙リストにおける形容詞の分布がもっともバランスがとれているといえるが、初級と中級の形容詞の語数の割合を調整する必要があるかどうか検討が必要である。日本語教育語彙表は、初・中・上級ごとのバランスがあるといえるが、中級後半と上級前半の形容詞を合わせた割合が全体の半分の量になっている。一方で上級後半の語数は少なく，語彙のレベル分けに偏りがある。

４．２　コーパスに現れる・現れない形容詞

語彙リストを比較するためには統一が必要である。統一は表記統一、品詞統一、形容詞の単位の統一であり、またそれによってコーパスの頻度数も調整する必要がある。たとえば、「すごい」と「凄い」は表記が違うため別の項目としてコーパスの語彙リストに現れるとき、その頻度数も再計算する必要がある。特にコーパスのデータを比較するに当たって、利用した形態素解析ツールおよび電子化辞書によって差異が見られる。たとえば、UniDic 短単位の特徴は、語彙の表記を統一した上、短単位で解析を行う傾向がある。IPADIC の特徴は、語彙の表記を別に捉え、複合語も単位として載せる傾向がある。本研究では、BCCWJ が利用している MeCab と UniDic の短単位のデータをベースにして、コーパスのデータを揃

えた。それで、JpWaC が ChaSen と IPADIC を利用したため形容詞のリストを UniDic で再解析した際、形容詞の語数が非常に変わったため、ある程度手で直した。手直しの差異、表記の統一をしたが、短単位で分けた複合形容詞をリストから亡くさないように元のまま複合の形容詞として保存した（たとえば、「興味深い」）。表 2 では、各コーパスにおける形容詞の語数が見られ、JpWaC の場合統一前と統一後のデータを示した。前述した既存の語彙リストにおける形容詞の語数と比較すると、大きな差異があることがすぐ見てとれる。

表 2　BCCWJ と JpWaC に現れる形容詞の語数

	BCCWJ-UniDic	**JpWaC-IPADIC[1]**	**JpWaC-再計算 UniDic-手直し[2]**
形容詞-一般	789	1522	903
接尾辞-形容詞的	12	7	0
形容詞-非自立可能	6	11	3
合計―形容詞―語数	807	1540	906

[1] 元々の ChaSen-IPADIC の品詞タッグは「形容詞―非自立」と「形容詞一接尾」。
[2] 再解析の後、5 頻度までのデータを手で直して、672 語になった。4 頻度以下は形容詞として分析されたデータだけを計算した（234 語）。

BCCWJ と JpWaC の語彙リストを比較した結果、二つのコーパスに現れる形容詞の分布は、ほとんど類似していることが分かった。両者ともあまり偏りがないデータと考えられ、２種のコーパス比較で得られたデータで他の既存のデータの評価ができると考えられる。両方のコーパスに現れる、あるいは現れない形容詞を観察すると、データ処理方法の差異しか見られなかった（いくつか残った表記問題を含む）。あるコーパスリストに無い語はコーパスに無いわけではなく、そのコーパスの処理方法の結果、取り出されていないケースが多かった。ここでは、とくにコーパス語彙頻度リストの作成にあたって、形態素解析の依存性またその問題点が見られる。BCCWJ リストにあるが、JpWaC リストにない、また JpWaC リストにあるが BCCWJ リストにない項目は、大ざっぱに言って、三つに分けられる。

- 短単位の形態素解析を利用したためおよび形態素解析の誤りで現れていない形容詞
- 表記の違いがある形容詞
- 低頻度で、限られた分野の形容詞である

JpWaC に無いが、BCCWJ にある形容詞は、学習者用の語彙リストに現れない、「けばい、労労じい、露けい」などの例があげられる。

一方、BCCWJ リストにないが、JpWaC にある形容詞は、複合形容詞か表記の違いのものであり、直接コーパスを文字列か違う表記で検索した結果、BCCWJ コーパスにも現れるものである。たとえば、「興味深い」のような複合の形容詞で、UniDic の短単位で二分以上の部分に分ける。たとえば、「興味深い」は「興味」と「深い」になる。同様の問題が見られるものとして，高頻度 100 語のみを対象にした場合、「興味深い、格好いい、詰まらない、勿体ない、数多い」がある。

４．３　語彙リストに現れる・現れない形容詞

前節に取り出した「興味深い、格好いい、詰まらない、勿体ない、数多い」の形容詞は語彙リストに扱われているか、どの難易度レベルで扱われているか調査した。結果を表 3

に示す。

表 3　短単位で取り出せなかった高頻度 100 語以内の形容詞が語彙リストにあるか

	国研国語語彙 DB	国研日本語語彙 DB	旧 JLPT 語彙リスト	TM 語彙リスト	日本語教育語彙表
興味深い	-	-	-	-	中級後半
格好いい	-	ある、級なし	-	-	-
詰まらない	1	-	4	-	-
勿体ない	1	1	2	-	中級前半
数多い	-	-	-	-	中級前半

国語教育語彙リストのデータベースでも三つの形容詞が無い理由は、一般的に複合語のデータが圧倒的に少ないこと、この語彙が近年頻度が高くなったことなどが考えられる。旧 JLPT 語彙リストは、三つ、日本語教育語彙表は二つの高頻度複合形容詞を扱っていないという結果が得られた。また，TM 語彙リストは UniDic 短単位を利用しているので、対象の形容詞は語彙リスト以外だと予想できる。「詰まらない」の語彙レベルは、統一されているが、「勿体ない」の場合、低学年のレベル(1)と中級レベルが見られる。以上の形容詞の頻度と分布から、それぞれの形容詞を学習項目として語彙リストに入れることが推薦できる。

さらに、BCCWJ の高頻度の形容詞の 100 語は、日本語教育語彙リストにカバーされているかを調べた。国研日本語語彙 DB、TM 語彙リスト、日本語教育語彙表では、100 語以内の形容詞がすべてカバーされている。しかし、表記統一には問題があるとよく見られる。旧 JLPT 語彙リストには、「幅広い」という形容詞がなかった。

高頻度の 100 語以降、特に 200 語以降の形容詞がリストにあるか無いかを検討すると、2 種の均衡コーパスおよびサブコーパスにおいて同じような分布を持っている形容詞のケースは多いが、散発的に数少ない形容詞がリストにはある。たとえば、「初々しい」は旧 JLPT 語彙リストと日本語教育語彙表にあるが、「歯痒い」が無い。同じように「心強い」と「名高い」は同じ分布なのに、前者だけが語彙リストに載っている。そのため、コーパスの頻度を基にして同じような分布の形容詞は新しい項目として日本語教育用の扱いを検討する必要がる。

高頻度の 200 語・300 語の形容詞のうち、日本語教育語彙表と旧 JLPT 語彙リストの中上級・上級のものがあるが、同じような特性を持った形容詞は扱っていない。逆に、その面から BCCWJ を基にした TM 語彙リストがもっとよくカバーされている。

４．４　クラスタ分析で見られる形容詞分布

コーパス頻度をもとに、高頻度語に対する統計的な分析を以下の手順で行った。1)BCCWJ の中で合計頻度 500 以上の形容詞 158 語を抽出。2)158 語に対する JpWaC の総頻度を抽出し、分析データを作成。3)BCCWJ の合計頻度やサブコーパスでの頻度、JpWaC の総頻度を対数変換。4) 対数変換済み値をもとに、SPSS で階層的クラスタ分析と主成分分析を行った。ただし、解析データに関して、一点だけ調整した。品詞や語彙の切り方の相違により、BCCWJ には形容詞として掲載されているが、JpWaC では形容詞として認定されていない 6 語「易い、旨い、らしい、〜ぽい、さり気無い、限り無い」は分析対象から外した。最終的には、152 語を対象に分析を行った。

階層的クラスタ分析におけるオプションとして、クラスタ法は、Ward 法を使用、サンプル間の距離定義は、ユークリッド距離を使用した。クラスタ分析の分類精度を評価するため、判別分析を行った。判別分析では、階層的クラスタ分析で出力したクラスタ数を従属

変数に、対数変換後の値を独立変数にして、変数同時投入法で解析を行った。判別分析の結果、6 個のクラスタの場合 88.2%、5 個のクラスタの場合 86.8%、4 個のクラスタの場合 90.1%、3 個のクラスタの場合 89.5%の予測精度が示された。この結果を受け、152 語のデータは、4 個のクラスタとして捉えるのがもっとも適切であると判断した。

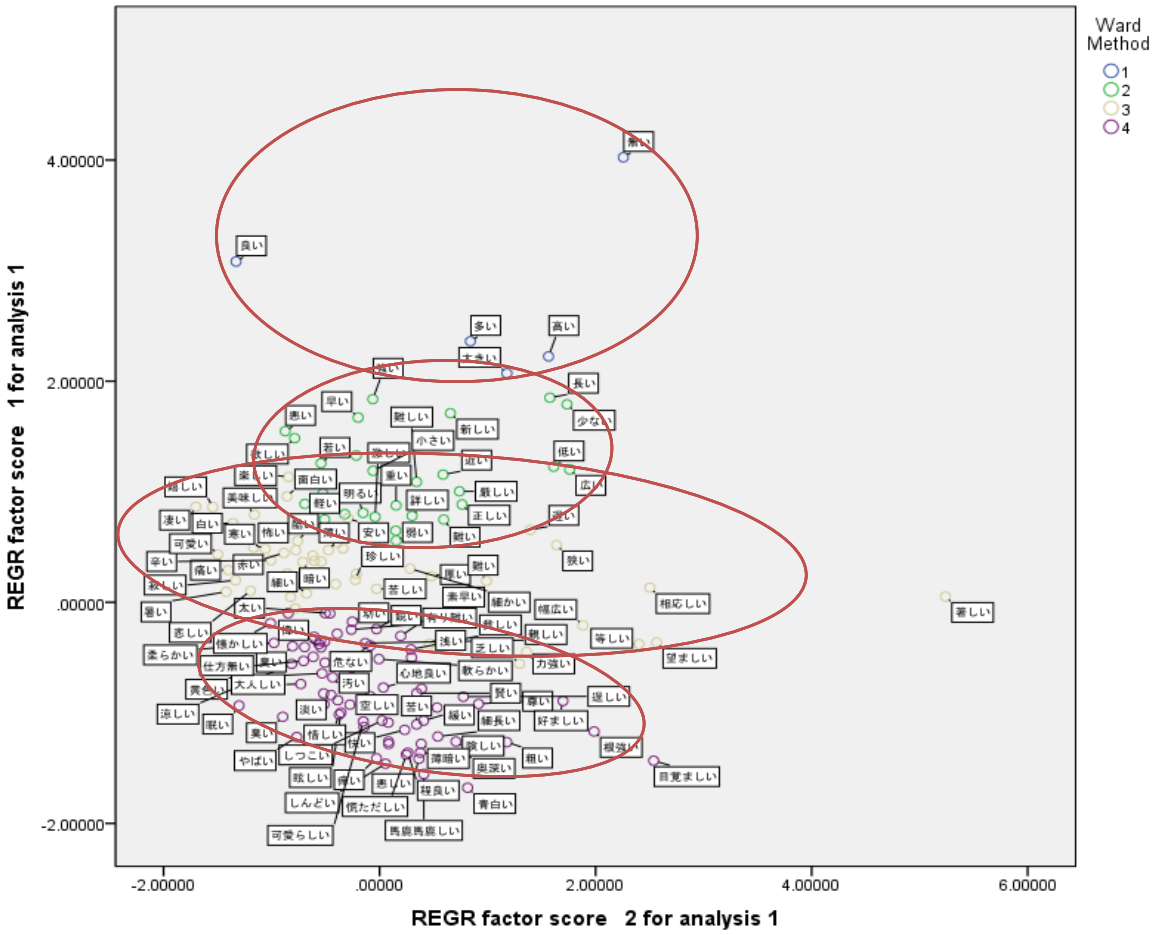

図2　第一主成分×第二主成分の得点によるサンプルの散布図

主成分分析では、クラスタ分析と同様に、対数変換済み値を使用した。第一主成分と第二主成分の合計固有値は、81.8%で、この二つの主成分で、8 割以上のデータが説明できる。また、Kaiser-Meyer-Olkin の標本妥当性の測度も 0.920 となり、説明力の高い分析であることが明らかになった。このことを踏まえ、クラスタ分析の結果と対応する形で、主成分得点をもとに、152 語の散布図を作成した。

図 2 に目立つ他のクラスタから離れている形容詞（無い、よい、著しい）は、特性を持っている語である。「無い」と「よい」は、非自立可能な形容詞であり、「著しい」は一番高頻度で特定目的白書のコーパスに現れ、偏りがある分布をもった形容詞である。既存の語彙リストでも、高い段階で教えている語である（JLPT:1 級、日本語教育語彙表：中上級、TM 語彙リスト：Inter）。

更に、各クラスタの具体例を表 4 に示す。

表4 クラスタの具体例

区分	ﾀｲﾌﾟ数	形容詞の例
ｸﾗｽﾀ1	5	無い、良い、多い、高い、大きい
ｸﾗｽﾀ2	27	弱い、短い、明るい、重い、激しい、強い、悪い、少ない、長い、早い、新しい、欲しい、深い、若い、古い、小さい、軽い、難しい、難い、正しい、低い、近い、優しい、広い、美しい、詳しい、厳しい
ｸﾗｽﾀ3	48	乏しい、力強い、浅い、等しい、幅広い、望ましい、相応しい、苦しい、厚い、著しい、恥ずかしい、青い、濃い、細い、凄い、悲しい、楽しい、細かい、美味しい、面白い、熱い、難い、嬉しい、暑い、冷たい、白い、忙しい、寂しい、可愛い、狭い、珍しい、酷い、安い、温かい、遠い、甘い、怖い、黒い、暗い、薄い、辛い、素晴らしい、痛い、可笑しい、寒い、赤い、固い、遅い
ｸﾗｽﾀ4	72	目覚ましい、馬鹿馬鹿しい、悪しい、青白い、根強い、重たい、奥深い、程良い、粗い、慌ただしい、しんどい、分厚い、険しい、醜い、尊い、快い、苦い、細長い、痒い、緩い、薄暗い、荒い、可愛らしい、眩しい、鈍い、切ない、逞しい、惜しい、空しい、好ましい、賢い、危うい、しつこい、心地良い、凄まじい、淡い、情けない、臭い、羨ましい、やばい、でかい、黄色い、辛い、涼しい、汚い、大人しい、軟らかい、貧しい、悔しい、危ない、怪しい、不味い、煩い、眠い、めでたい、素早い、臭い、柔らかい、丸い、親しい、きつい、幼い、久しい、懐かしい、有り難い、物凄い、鋭い、偉い、恐ろしい、太い、仕方無い、宜しい

各クラスタの解釈のため、BCCWJ の総出現頻度をもとに、平均値を確認した。クラスタ 1 は、平均頻度が 161732.2 となり、超高頻度の形容詞である。クラスタ 2 は平均頻度が 13993.4 となり、高頻度の形容詞と言える。クラスタ 3 は 5673.4 となり、一定量の使用が確認されるが、高頻度で初級でもよく教える語もある。クラスタ 4 は、平均頻度が 1523.9 となり、比較的に頻度も低く、難易度も高い語彙が多いことが確認された。クラスタ 1、2 は初級学習者には必須、クラスタ 3 は、初級・中級学習者に分けられ、クラスタ 4 は主に中・上級学習者向けの形容詞であるが、「黄色い」および「円い」のような普段初級で学習される語彙も現れる。クラスタで得られた結果は、サブコーパスおよびコーパスごとの頻度・分布を基にしたグルーピングで、直接語彙習得段階と結びつけにくいところもあり、他の要因を考慮に入れつつ教育のために利用可能である。また、同じような方法で 158 語以外の形容詞も分析する必要がある。

５．新しい形容詞の語彙リストに向けて

既存の語彙リストとコーパスを分析した結果、新しい形容詞の語彙リストを作成するメリットがあることが明らかになった。本研究で収集した形容詞のデータを更にデータベース化して、統一し、比較できるような表として提供することが望ましい。今回は BCCWJ およびそのサブコーパスによる形容詞の語彙頻度のリストをベースにしたが、今後、今月公開された超大規模 JpTenTen ウェブコーパスのデータを利用する予定である。また、今回の研究に UniDic の短単位を利用したが、長単位のデータが抽出できるようになったので、複合形容詞を適切に扱うために長単位にデータを揃えてデータベース化する予定である。2 節に取り上げた語彙リストにあるデータを統合し、形容詞の見出し語以外に、表記、品詞、それぞれの語彙リストにあるかどうか、その取り出した難易度レベル、コーパス・サブコーパスごとの頻度、形容詞の形式、意味分類などの情報を提供することを目標としている。

６．まとめと今後の課題

本論文では、形容詞を対象にして既存の日本語の様々な語彙リストと大規模な 2 種のコーパスから取り出した頻度リストを比較した。その結果、形容詞の語数およびカバーされた形容詞、その難易度レベルの間にキャップがあると判断された。例として取り出した既存の語彙リストにはない高頻度の形容詞は、今後の語彙リストの項目として入れる必要が

ある。

どの既存の語彙リストも全体の大規模な現代日本語均衡コーパスか大規模な現代日本語の調査を利用していないため、語彙リストの再作成が必要であると考えられる。また、コーパスデータを用いた語彙リストは、形態素解析および電子辞書の言語処理方法への依存性があるということが確認され、特に複合形容詞のデータはほとんどカバーされていない。そのため、今後 UniDic の長単位の分析結果が望まれるとことである。

謝　辞

本研究は、博報財団第 7 回「日本語海外研究者招聘事業」「日本語教育における語の共起関係」という研究（平成 24～25 年度、招聘研究員：スルダノヴィッチ・イレーナ）および「研究種目と分野：基盤研究(A)日本語教育研究課題名：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（研究代表者：砂川 有里子（筑波大学)」による支援を得ています。「教育基本語彙の基本的研究」のデータベースおよび BCCWJ のコーパスを含めて国立国語研究所が研究環境を与えてくださったことに感謝いたします。

文　献

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, pp. 275-285.

国際交流基金・(財)日本国際教育協会（1994）『日本語能力試験出題基準』凡人社

教育基本語彙の基本的研究―増補改訂版―（2008）国立国語研究所報告 127, 明治書院

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

スルダノヴィッチ・イレーナ, 仁科喜久子（2008）「コーパス検索ツール Sketch Engine の日本語版とその利用方法」『日本語科学』23 号, 国書刊行会, pp. 59-80

スルダノヴィッチ・イレーナ（2012）「複数のデータを活用したイ形容詞と名詞のコロケーションの記述―日本語教育のための資料作成を目指して―」第 82 回 NINJAL サロン, 2012 年 11 月 27 日

大学英語教育学会基本語改定委員会(編) (2003)「大学英語教育学会基本語リスト: JACET List of 8000 Basic Words」大学英語教育学会

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（特集 :「語彙の教育」)』27-10, 明治書院, pp. 50-58.

松下達彦 (2010) 「日本語を読むために必要な語彙とは？ ー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」『2010 年度日本語教育学会春季大会予稿集』pp. 335-336.

松下達彦 (2011) 「日本語を読むための語彙データベース」(The Vocabulary Database for Reading Japanese) Ver. 4.0. (http://www.geocities.jp/tatsum2003/よりダウンロード可能)

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

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

Nation, Paul (2001) Learning vocabulary in another language. Cambridge University Press

Srdanović, Irena, Bekeš, Andrej, 仁科喜久子(2008)「複数のコーパスに見られる副詞と文末モダリティの遠隔共起関係」特定領域研究,「日本語コーパス」平成 19 年度公開ワークショップ（研究成果発表会）予稿集, pp. 223-230.

関連 URL

国立国語研究所の言語コーパス整備計画 KOTONOHA　http://www.ninjal.ac.jp/kotonoha/
日本語教育語彙表の検索システム「学習項目解析システム」http://lias.intersc.tsukuba.ac.jp/
中納言検索システム（BCCWJ）https://chunagon.ninjal.ac.jp/
スケッチエンジン検索システム（JpWac, JpTenTen）http://www.sketchengine.co.uk/
TM 語彙リスト http://www.geocities.jp/tatsum2003/
形態素解析辞書 UniDic http://download.unidic.org/twitter.com/unidic