国立国語研究所学術情報リポジトリ

大規模コーパスを用いた形容詞と名詞のコロケーションの記述的研究：日本語教育のための辞書作成に向けて

メタデータ	言語: Japanese 出版者: 公開日: 2015-10-30 キーワード (Ja): キーワード (En): collocation, corpus, adjectives+nouns, Japanese language education, Japanese linguistics 作成者: スルダノヴィッチ, イレーナ, SRDANOVIC, Irena メールアドレス: 所属:
URL	https://doi.org/10.15084/00000515

国立国語研究所論集 (NINJAL Research Papers) 6: 135–161 (2013)
ISSN: 2186-134X print/2186-1358 online

大規模コーパスを用いた形容詞と名詞のコロケーションの記述的研究
――日本語教育のための辞書作成に向けて――

イレーナ・スルダノヴィッチ

リュブリャーナ大学／国立国語研究所 外来研究員［–2013.09］

要旨

近年，日本語のコロケーション辞典など，コロケーションを記載したリソースも現れてきたが，現代日本語の大規模コーパスを用いた記述的コロケーションデータはまだない。また，直感と経験に基づいて作成された日本語教科書などの教育用の教材においても，コロケーションに関しては注目度が低い。そこで本稿では，「形容詞＋名詞」の組み合わせによるコロケーションに焦点を当て，BCCWJ・JpTenTen という 2 つの現代日本語コーパスからコロケーションを取り出し，1)「形容詞と名詞のコロケーションデータ」，2)「日本語教育のための形容詞と名詞のコロケーション辞書」の 2 種のリソースの作成方法を提示し，「高い」を記述モデルの一例として日本語教育への応用方法を示すことを目的とする。1）の「形容詞と名詞のコロケーションデータ」は，500 語の形容詞を対象にして，シンタクスを考慮に入れて抽出した名詞とのコロケーションおよびその前後文脈をコーパスごとに整理し，比較できるようにするものである。現時点では，100 億語のコーパス JpTenTen から取り出した 500 語の形容詞とその名詞とのコロケーションデータ（23247 語）を取り出すことができ，BCCWJ からの抽出は進行中である。2）の「日本語教育のための形容詞と名詞のコロケーション辞書」は，すべての形容詞の 62% をカバーする 25 語の基本的な形容詞について詳細に記述することを目指す。そこで，高頻度の形容詞「高い」を取り上げ，コロケーションデータの分析結果を提示し，前述の「形容詞と名詞のコロケーションデータ」を基にした「日本語教育のための形容詞と名詞のコロケーション辞書」の基盤作りを示す。能力レベルによって分類された辞書項目は，被修飾名詞の語彙マップを作成したり，ジャンルごとの特有な情報を併記したりして，学習者の学習困難なコロケーションに焦点を当てて記述する。最後に，これらのデータが示唆する様々な理論的･応用的研究の発展可能性について検討する。このような形容詞のコロケーションデータが整備されることにより，従来，日本語を対象としては作成されてこなかったデータを提供し，今後の日本語学の語彙と文法の研究や資料作成，および日本語教育用教材・シラバス作成のために資することが期待できる *。

キーワード：コロケーション，コーパス，形容詞＋名詞，日本語教育，日本語学

1. はじめに

大規模コーパスと検索ツールの構築とともに，言語現象の実証的・記述的な研究にも新しい局面が開かれつつある。そのうち，パターン化されている語と語・語と節などの言語要素の組み合わせの存在, すなわちコロケーションの用法および制限に関する情報の重要性が強調されてきた。

従来の構造言語学，生成文法などの方法論では，単語，文法規則，文の構成に関する研究が盛んに行われてきたが，複数単位（MWU: Multi-Word Unit）および節のようにまとまった範囲に

* 本稿は，博報財団第 7 回「日本語海外研究者招聘事業」の招聘研究員として国立国語研究所に滞在している間（平成 24 年 10 月～ 25 年 9 月）の進行中の研究の成果をまとめている。データは来年度中に公開する予定である。

関する実証的データが不十分であった。このような研究領域は，コーパス言語学と共に新しく開発されてきた統計的な面から見たコロケーション，または以前からロシアや東欧で研究されてきた慣用連語の面から見たコロケーション，およびその後現れた認知言語学の面から見たコロケーションの研究にとって，研究の余地を残してきたと言える。その後，コーパス構築とコーパス分析方法の進歩に伴って，試行錯誤の中でコロケーション研究も徐々に進められてきた。この状況の中で，コロケーションの定義の曖昧さや確実な統計的方法論が確立されていないなどの問題が残されているものの，限られた 5 単語以内のコロケーションスパン，または，2 単語に絞ったコロケーション抽出などのアプローチは，コロケーション研究においてなお影響を与える。

言語学・第二言語教育などの分野において，コロケーション研究で発見されたことは，確実に大きな利点があった。例えば，語彙の組み合わせには，ある状況下で多くの人が繰り返す言語パターンが見られる，ということの実証的な発見である（McCarthy & Carter 2006）。ある語とある語が組み合わせられる確率と，そこにかかる制限，という 2 つの側面から捉えることができる。ある組み合わせの確率が高いということは，母語話者がよく利用する表現であることを示している。また，ある組み合わせに制約があるということは，母語話者がその表現を自然に発話するか否かという点に関わるものである。これらの 2 側面は，言語表現としての自然さと，言語コミュニケーション上の自然さという点に関連していることである。

母語話者は通常，コロケーションの確率と制限を意識しないで正しい組み合わせを産出し，うまく言語コミュニケーションを行うが，自分の直感や内省でコロケーションについての体系的な情報を取り出すことには極めて困難が伴う。第二言語学習者が理解・産出するコロケーションに関する研究によると，目標言語のコロケーションは誤った表現や不自然な表現の原因になりやすいということが明らかにされている（James 1998, Nation 2001）。学習者がコロケーション情報の学習を効果的に学習すべきこと，および，外国語教育においてコロケーションを体系的に導入すべきことが指摘されている。本研究で扱う実証的方法論の大きな利点は，従来の内省に基づく方法では明らかにできなかった言語的情報を大規模なデータから取り出し，日本語学と日本語教育に応用できるコロケーションデータを提供することである。

本稿では，言語学におけるコロケーションの位置付けについて述べた上で，日本語のコロケーション研究の背景および日本語のコロケーションデータの重要性について述べる。続いて，形容詞と名詞を対象にして大規模コーパス（BCCWJ）と超大規模コーパス（JpTenTen）を用いた日本語のコロケーションの抽出・分析・記述について述べ，2 種のリソースの作成について進行中の研究成果を紹介する。第一のリソースは「形容詞と名詞のコロケーションデータ」で，500 語の形容詞を対象にし，シンタクスを考慮に入れた 2 種のコーパスからのコロケーションデータの抽出方法，比較方法および記述を提示する。現時点では，500 語の形容詞とその名詞とのコロケーションデータを JpTenTen から取り出すことができ，BCCWJ からの抽出は進行中である。第二のリソースは「日本語教育のための形容詞と名詞のコロケーション辞書」で，前述の「形容詞と名詞のコロケーションデータ」を基に，高頻度の形容詞「高い」を取り上げて，難易度，予想しやすさなどの要因を考慮に入れた日本語教育のための名詞とのコロケーション辞書作成の基盤作

りを示すことである。最後に，作成資料を基に，理論の面および応用の面から得られる示唆について論じる。

2. 研究背景

2.1 言語学におけるコロケーションの位置付け

コロケーション研究の始まりは，記述的研究，言語教育に向けた辞書学（斎藤 1907, Palmer 1938, Hornby 1954），文体研究（Firth 1951, Yamamoto 1958, McIntosh 1966）において見られる。連語，コロケーション，共起などの用語が使われてきたが，いずれにしても，その研究は実証的なアプローチをとり，広義での機能主義論の言語観と一致し，言語の機能的，運用的側面を重視するものである。

コンピュータを用いたコロケーション研究は，特にロンドン派の Firth（1951）のアイディアに影響を受け，その初期においては，辞書作成への応用を念頭に，語の意味に焦点を当ててきた（Sinclair 1966）。同時に，Halliday（1966）は，選択体系機能文法の枠組みの中で，より広い文脈を考慮にいれた文法的な語彙研究をコロケーション研究として実践した。その後も，語彙意味論を中心にした新ファース派（Neo-Firthian）による言語研究の中で，コロケーションを対象としたコーパス研究が進んできた。この方法論は，主に単語を抜き出し分析するコンコーダンサーによるデータを検討しつつ，繰り返すパターンとその頻度によってコロケーションの重要性を訴えるものである。一方，統計的な面から見たコロケーション研究は，共起関係の傾向を統計的に計算するための統計値を何種類か提案している（Hunston 2002）。このアプローチによって，語と語の間のスパンを 3 ～ 5 語に限定し，単語と単語の組み合わせの強さを計算する。

近年では，シンタクスを組み入れて考える方法，およびシンタクスを超えて考える方法が見られるようになってきた（Grefenstette 1992, Stefanowitsch & Gries 2003）。統計的研究に統語的アプローチを加えることによって，コロケーションデータの取り出し方をさらに精密なものにすることができるようになる。シンタクスや統計的アプローチを超えた研究としては，Hoey（2005）が挙げられる。それによると，コロケーションの発生は心理的に起こるものであり，内側にある心のあり方と関連しているという。

コンピュータを用いないコロケーションの研究としては，ロシア・東欧の慣用表現，およびドイツで行われている結合価（動詞が必要とする格関係）の研究が見られる。それらの影響を受けた日本語研究については，田野村（2012），堀編（2012），荻野・荻野（2007）を参照されたい。

2.2 コロケーション習得の重要性

単に語と文法規則を習得しただけでは，学習者が自然な表現を作ることは難しい，ということは，先行研究でよく指摘されている。学習者にとっては，作文をするときだけでなく，文章を理解する場合においても，複合単位による表現習得は不可欠である（McCarthy & Carter 2006）。

Nation（2001）によると，コロケーションは，しばしば文法的・語彙的に予測不可能（unpredictable）であるため，学習者にとっては誤りやすく，ネイティブらしい自然な表現になりにくいものであ

る。コロケーションの予測しにくさは，学習する言語と学習者の母語および他言語におけるコロケーションの違いに起因する。例えば，英語の「to make a tea」は，日本語に直訳すると「お茶を作る」となり，正しい「お茶を入れる」という表現に対し「お茶を作る」は不自然になる。同じように，cold という形容詞は「冷たい」「寒い」という 2 つの意味を持つので，「cold water」というコロケーションは，「冷たい水」ではなく「寒い水」という誤訳を引き起こしやすくなる。

第二言語学習におけるコロケーション研究は，学習者コーパスを用いたコロケーションの誤りの検討，母語話者と学習者のコロケーションの違いなどを扱っている。Kjellmer（1991）は，自然で母語話者らしい発話ができるようになるために，コロケーションは重要な学習項目であると述べている。また，語彙項目は単独で教えるのではなく，個々の語からコロケーションに重点を移すことが必要であると主張している。James（1998）も「自然さ」のためのコロケーション習得の重要性を指摘している。母語話者と学習者のコーパスを比較した研究結果によると，学習者が利用しているコロケーション情報の中には，過剰に利用するコロケーションと十分利用されていないコロケーションが見られ，学習者は母語話者のようにコロケーションを使いこなすのが難しいことが指摘されている。コロケーションの学習の負担の大きさは，母語または既習の第二言語のコロケーションから，そのコロケーションが予想できるかどうかに依存する。言語学習でコロケーションを導入するに当たっては，特にどのコロケーションが高頻度で現れるか，および高頻度の語彙項目が，予測不可能なコロケーションとしてどのように出現するかを検討する必要がある。

日本語の形容詞からなるコロケーションの誤用を分析した研究の例としては，家田（2003），曹・仁科（2006），曹（2012）などが挙げられる。曹・仁科（2006）では，誤用には母語に起因するものが多く，発達的誤用・過剰一般化の誤用などがあると述べている。曹（2012）では，中国語母語話者の作文ではコロケーションの誤用が20%近くになっていることが報告されており，誤用が生じる表現は一部の初級レベルの属性形容詞に集中しているという。誤用の理由は，形容詞の意味の不十分な理解，統語的な制限による誤用，存在しない組み合わせによる誤用などが挙げられている。同じ論文では，フランス人の作文のデータでも直訳による誤用が確認され，意味の誤りとして日本語教育でいう 4 級の属性形容詞のコロケーションが多く見られると指摘されている。

2.3 従来の日本語コロケーション研究および今後の必要性

本節では，従来の日本語コロケーションリソースに関する研究成果を紹介する。その上で，今後どのようなリソースを作成するのが望ましいかについて述べる。

1980 年代に開始された EDR 日本電子化辞書プロジェクト（EDR: Japanese Electronic Dictionary Research Institute）によって，EDR 辞書，コーパスなどが作成され，言語処理の共通データとして広く用いられるようになった。計算機用日本語基本辞書 IPAL（IPAL: Information-technology Promotion Agency Lexicon）は，1980 ～ 1990 年代に作成された辞書で，基本和語動詞 861 語，基本形容詞 136 語，文法的に特徴のある名詞 1081 語が収録されている。

コロケーション辞典には，EDR に基づいた「名詞・格助詞・動詞」を対象にした『日本語動詞の結合価』（荻野他 2003），分野別のコロケーションを集めた『知っておきたい日本語コロケーション辞典』（金田一 2006）などがある。『日本語表現活用辞典』（姫野 2004）は，最初の学習者用のコロケーション辞典であり，中級学習者用文芸作品，新聞，他の辞典などのデータ を利用し，動詞，形容動詞のコロケーションを記述したものである。その大幅増補の改訂版としては，動詞，形容詞，形容動詞をカバーしている『研究社日本語コロケーション辞典』（姫野監 2012）が出版された。『てにをは辞典』（小内 2010）は，20 年かけて採集した結合語の辞典で，250 名の作家の作品から語と語の結びつき 60 万例を採録した辞典である。なお，日本語コロケーションに関する研究（荻野 2008，田野村 2010，2012，茂木 2012，スルダノヴィッチ 2012），コロケーションに関する学習者用の教材（秋元・有賀 1996，小野他 2010）などが挙げられる。

コーパスからコロケーションを抽出するツールとしては，「茶漉」（深田 2007），日本語のための「スケッチエンジン」（Srdanović et al. 2008，スルダノヴィッチ・仁科 2008，スルダノヴィッチ他 2013），「なつめ」（仁科監 2012），「NINJAL-LWP」（プラシャント・赤瀬 2012）などが見られる。また，16 億文のウェブテキストに基づいて格フレーム [1] を自動的に取得した研究がある（河原・黒橋 2006）。

以上の研究成果を見ると，特にこの 10 年間は多くのコロケーション研究が行われてきたことが分かるが，以下のような観点から今後なお発展の余地があると言える。

- 現代日本語の大規模コーパス，およびバランスが取れたコーパスを基にしたコロケーション辞書，リストなどのデータが望ましい。従来のコロケーション辞典は，近代･現代のデータが混在しているものがある。
- 複数の現代語のコーパス，およびサブコーパスを利用したジャンル別のコロケーションの考察にはメリットがある。
- 日本語教育用の日本語コロケーションの辞典などの教材は，上述の現代語のコーパスを利用することで，学習者のために予想しにくいデータを強調し，難易度別，ジャンル別，語彙の意味分類を付すことにより，さらに有用になると考えられる。
- コロケーション抽出ツールは，何種かのコーパスから様々なコロケーションタイプを取り出せるが，コロケーションタイプをさらに増やすこと，および整理することが望ましい。この際，コロケーションの前後文脈を考慮に入れながら，コロケーションが現れている構文を検討し，2 単位以上を取り出す必要がある。

2.4 対象にした形容詞と名詞のコロケーション

前節で示したように，現時点においては，大規模な現代日本語コーパスを用いたコロケーションリスト・辞典は存在しない。それは，言語研究の観点から見れば，現代日本語のコロケーショ

[1] 用言とそれに関係する名詞を用言の各用法ごとに整理したものである。

ンが言語的情報として十分には記述されていないことを意味する。日本語教育の観点から見ても，学習者が利用できる現代日本語のコロケーションに関する情報が不足しているということになる。そこで本研究では形容詞と名詞の組み合わせに絞り，コロケーションデータの記述的研究を行う。これにより，日本語学における文法や語彙研究，または，より応用的に日本語教育に役立てることを目的とする。

コロケーションとしての組み合わせには多様な種類が考えられるが，ここではプロジェクトの時間的な制約により，研究対象を形容詞と名詞の組み合わせに絞った。従来の日本語コロケーション研究は，述語の格フレームに基づく共起関係の記述に偏っており，形容詞と名詞の組み合わせについてのデータが非常に少ないという事実が，本研究を実施する 1 つのきっかけとなっている。例えば，コロケーション研究の例としての格フレームのデータ（河原･黒橋 2006）は，「名詞＋助詞＋動詞」[2] のデータを提供し，形容詞とのコロケーションは「形容詞＋動詞」になっている。また，前述したコロケーション辞典（荻野他 2003，金田一 2006）は「名詞＋助詞＋動詞」のデータに絞っている。形容詞と名詞の組み合わせは，最初の学習者用のコロケーション辞典（姫野 2004）でも対象になっていなかったが，最近の近代･現代日本語のデータを基にした辞典（姫野監 2012，小内 2010）においては扱うようになってきた。形容詞の連体の用法，および形容詞と他の語の組み合わせのパターンに関しての研究の重要性は，従来の研究でも認められたことがある。特に IPAL 辞書と関連した研究に繋がり（宮島 1993，橋本・青山 1992），また形容詞の記述的研究にも見られる（西尾 1972，八亀 2008）。現代日本語の大規模コーパスが利用できる現時点でも，このような量的・記述的研究を深めて，続ける意義がある。

なお，本稿では，形容詞を見出し語にして，それと共起する名詞のコロケーションデータを示す。日本語学習者を利用者として考えると，コロケーションデータには形容詞を見出し語にするメリットが様々ある。そのうちのいくつかの例を挙げると，①それぞれの形容詞と結びつく様々な名詞のコロケーションから，その形容詞の用法・意味について体系的に知ることができる，②選んだ形容詞と結びつく名詞の意味セットを語彙マップから知ることができ，名詞と形容詞の学習が効率的にできる，③形容詞は「義務的ではない」ものがあり，産出しなくてもコミュニケーションができるため，形容詞の習得と産出は遅くなる場合がある。そこで，形容詞の習得を支援できるデータを提供する。コロケーションから見出し語の用法・意味を把握できることについては Kilgarriff & Rundell（2002）を，意味に基づいてマッピングした語彙マップを用いた学習の効果については Morin & Goebel（2001）を，形容詞の義務的・義務的ではない利用については，八亀（2008）をそれぞれ参照されたい。

一方，名詞を見出し語にした，それと共起する形容詞などの品詞についてのコロケーションデータも必要である。先行研究でも，両方の必要性について言及されている（橋本 2007）。既に指摘しているように，学習者のコロケーション習得は大きな課題であり，今後できる限り様々な方向

[2] 自然言語処理研究および日本語の文法研究において，結合価による動詞の研究，いわゆる動詞の格支配についての研究が多数ある。その詳細については荻野・荻野（2007）を参照されたい。

から体系的なコロケーションの扱いを教育に導入する必要がある。

第二言語教育において，高頻度のデータを取り上げるメリットについては，多数の指摘がある。最も重要な点は，大規模なデータの分析によると，高頻度の 2500 語はデータの 70 ～ 80% をカバーしていることで，他の要因を考慮に入れつつ，学習者に最初に教えるのが効率的である（Nation 2001，松下 2011）。本稿では，まず高頻度のデータを示し，その上で，日本語教育に幅広く利用されている難易度レベルの情報および学習負担と関連する，他言語から見てのコロケーションの予想しやすさを取り入れる。

なお，日本語でいう形容詞というカテゴリーは普遍的ではなく，言語ごとにバリエーションが多いので，他言語に形容詞として現れる語彙から見たコロケーション情報も今後の課題として必要である。例えば，英語の「old man」は，日本語では「老人」という名詞として使われており，「old」の意味に対応する日本語の形容詞「古い」は使われていない。このような，学習者にとって重要な情報をカバーするために，他言語の母語話者による形容詞の産出からデータを見る必要がある。それに当たって，形容詞だけではなく，形容動詞（例：元気な），複合名詞（例：青信号）などの記述的な機能を持つ単位もあわせて把握する必要がある。

3. 利用するコーパス，ツールなどのリソース

形容詞と名詞のコロケーションを取り出し，分析するために，現代日本語の大規模な書き言葉コーパスを利用する。第一に，現代日本語の書き言葉をバランスよく収録した 1 億語の「現代日本語書き言葉均衡コーパス」（以下 BCCWJ）を利用する。第二に，100 億語の超大規模日本語ウェブコーパス（以下 JpTenTen）を利用する [3]。

3.1「現代日本語書き言葉均衡コーパス」（BCCWJ）と中納言

「現代日本語書き言葉均衡コーパス」（BCCWJ: Balanced Corpus of Contemporary Written Japanese）は，現代日本語の書き言葉を対象として，母集団に対する統計的な代表性を有するサンプリングが実施されており，日本語では初となる均衡コーパスとして設計された（前川 2008，山崎 2009，丸山他 2011）。1 億語超のデータは，複数のサブコーパス [4] によって構成されており，形態素解析ツール MeCab および電子化辞書 UniDic を用いた形態論情報がアノテーションされている（小木曽・伝 2011）。さらに，UniDic を用いた「短単位」を，長単位解析器 Comainu（小澤他 2011）により組み上げる形で「長単位」のアノテーションが施されている。なお，コーパス検索ツールとしては，BCCWJ を対象に開発されたウェブ上の「中納言」[5] を利用する。中納言は，短単位・長単位・文字列の 3 つの方法によって検索を実施することができる。

[3] 本研究が始まった時点では，JpWaC（Srdanović et al. 2008）と BCCWJ を比較し，予備的研究ではコロケーションデータを分析した。その後，サイズがさらに巨大で，アノテーションの面でも BCCWJ に準拠している，新しく構築した JpTenTen を利用するようになった。予備的研究で既に明らかになったように，2 種のコーパスの利用は，いくつかの点でメリットがある。それについて，さらに 4 節で説明する。
[4] 詳細は，http://www.ninjal.ac.jp/corpus_center/bccwj/doc/manual/BCCWJ_Manual_02.pdf を参照されたい。
[5] https://chunagon.ninjal.ac.jp/

3.2 超大規模ウェブコーパス「JpTenTen」とスケッチエンジン

100 億語の日本語超大規模コーパス JpTenTen は，TenTen 群の 1 つとして 2011 年に構築され（Pomikálek & Suchomel 2012），スケッチエンジン（Sketch Engine）というレキシカルプロファイリングのツール[6]に搭載された（スルダノヴィッチ他 2013）。JpTenTen の構築は，日本語のウィキペディアを利用し，日本語言語モデル学習を行ったもので，ウェブ上のテキストデータをクロールし，様々な方法やツールでクリーニングを行った。このコーパスは，BCCWJ と同じように MeCab および UniDic を利用して形態素解析されている。コーパス全体は短単位として処理されているが，長単位を用いた 2 億語のサンプルコーパスがある。JpTenTen の構築手段，アノテーションおよびそのレキシカルプロファイリングについては，スルダノヴィッチ他（2013）を参照されたい。

スケッチエンジン[7]は，ウェブ上のコーパス検索ツールであり（Kilgarriff et al. 2004），レキシカルプロファイリング手法を用いた共起・文法関係の分析機能を持つ。これは，キーワードの複数のコロケーション・文法関係の情報をまとめた形で短時間に抽出し，表示する「ワードスケッチ」という機能である。日本語のコロケーションデータを取り出すために，日本語の「文法関係ファイル」（正規表現，品詞，活用形，単語から取り立てた規則の設定ファイル）を作成した（Srdanović et al. 2008，スルダノヴィッチ他 2013）。

3.3 その他のコーパスの利用

主に利用する BCCWJ と JpTenTen 以外に，必要に応じて，他のコーパスを参考にする。例えば，話し言葉におけるコロケーションの傾向を確認するために，インフォーマルな会話データの「名古屋大学会話コーパス」，フォーマルな会話データの「及川コーパス」などを利用する。今回の研究の主な対象ではないが，他言語におけるコロケーションの用法を検討するために，他言語のコーパスも利用する（例えば，英語の BNC とウェブコーパス，スロベニア語の GIGAFida，セルビア語のウェブコーパスなど）。対訳コーパスとしては，和英対訳コーパス，和スロベニア語の対訳コーパスなどが挙げられる。他言語のコーパスは，他言語の母語話者にとって，日本語のコロケーションが予想しやすいかどうかを確認するために利用できる。

さらに，コロケーションの誤用を検討するために，学習者コーパスを利用する。例えば，「寺村誤用例集データベース」[8]，「日本語学習者の縦断的会話コーパス」[9]（迫田他 2012），「オンライン日本語誤用辞典」[10]，「ナツメグ支援システム」[11]（仁科監 2012）などである。

---

[6] レキシカルプロファイリング手法とは，コーパスから抽出したコロケーションや文法的振る舞いなどを網羅的に提示することである。
[7] https://the.sketchengine.co.uk
[8] http://teramuradb.ninjal.ac.jp/
[9] http://c-jas.jpn.org/
[10] http://cblle.tufs.ac.jp/llc/ja_wrong/index.php?m=default
[11] http://hinoki.ryu.titech.ac.jp/nutmeg/

3.4 難易度の判定リソース

本研究では，『日本語能力試験出題基準』（1994）（改訂版は国際交流基金 2002）の旧語彙リスト（以下「旧 JLPT 語彙リスト」）をそれぞれの形容詞およびそのコロケーションを構成している単位の難易度の判定リソースとして使う。この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最も幅広く利用されており，その影響が現在よく用いられている日本語教科書にも見られる。語彙難易度は，「学習到達度を測定する」という日本語教育の観点から 4 段階に分かれ，下位の 4 級から上位の 1 級までである。この語彙リストでの語数は 1 万語ぐらいで，それ以外の語彙は級外と判定する。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，非公開になっているため，本研究では利用できない。

3.5 語彙意味セットを判定できるリソース

コーパスから得られた各形容詞と共起する複数の名詞のデータを意味セットに分けて，名詞の語彙マップのモデルを作成する。それに主に利用するリソースは，既に形容詞の様々な語義を認知言語学の観点，語彙意味論の観点から記述している以下の文献である。

・八亀裕美（2008）『日本語形容詞の記述的研究―類型論的視点から』明治書院
・西尾寅弥（1972）『形容詞の意味・用法の記述的研究』秀英出版
・今井新悟（2011）『日本語多義語学習辞典　形容詞・副詞編』アルク
・国立国語研究所編（2004）『分類語彙表　増補改訂版』大日本図書

4. 形容詞と名詞のコロケーションデータの作成方法

本節では，「形容詞＋名詞のコロケーションデータ」および「日本語教育のための形容詞と名詞のコロケーション辞書」の作成手順および方法について述べる。

4.1 形容詞の語彙リストの作成と形容詞の見出し語の決定

予備的分析として，形容詞を対象にして既存の日本語教育用の様々な語彙リスト同士を比較するとともに，それらのリストと大規模な 2 種のコーパスから取り出した頻度リストとを比較した（スルダノヴィッチ･李 2013）。その結果，各語彙リストにカバーされている形容詞の語数，種類，難易度レベルの間にギャップがあることが明らかになった。高頻度の最初の 100 語の形容詞は，いくつかの語の複合形容詞をのぞけば，すべてのリストでカバーされているが，100 語～ 200 語の高頻度の形容詞のカバー率はリストによって異なる。どの既存の語彙リストも大規模な現代日本語均衡コーパスや大規模な現代日本語の調査を参考にしていないため，語彙リストの再編成が必要なことが明らかになった。なお，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，長単位のデータも語彙リストに入れ，長単位と扱うべきだがそう扱っていない形容詞の検討を行った（Srdanović 2013）。現在，この研究の成果の 1 つと

して，BCCWJ と JpTenTen のデータ[12] を用いた新しい形容詞の語彙リストを作成し，公開するために整理中である。その内容は，コーパスから抽出した形容詞，その頻度，他の語彙リストとの比較などである。

さらに，大規模コーパスではサブコーパス別，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っており，英語では高頻度の 2000 の基本語彙がテキストの内容の 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もある（Nation 2001）。日本語では，松下（2011）が Nation と同じ枠組みで BCCWJ のモニター版を用いて「TM 語彙リスト」[13] を作成している。このリストは，一般用のデータとしては基本 2500 語を含み，総計 20326 語である。基本の 2500 語のうち，形容詞は 93 語である。この形容詞は最も基本的な高頻度語彙であり，様々なサブコーパスに現れるため早い段階で導入して学習させることが提言されている。Srdanović（2013）によれば，BCCWJ の長単位の形容詞のデータを分析した結果，高頻度順で 25 位までが 62%，127 位までが 90% 程度の形容詞の利用をカバーできているので，これらの形容詞を学習者に優先的に教えるべきだと考えられる。次の中頻度の 341 語の形容詞は 7% 程度をカバーする。低頻度の形容詞は，長単位の異なり語数で見ると，カバー率が非常に高く（96%），複合辞成分としての形容詞の造語性は非常に高いということが明らかになった。「っぽい，らしい，ない，やすい，深い，臭い」などの接尾辞を用いると，数多くの複合形容詞が形成されるので，学習者がそのような語の産出能力を付けることは有意義であると言える。例えば，「臭い」からなっている複合形容詞は，長単位の BCCWJ において 280 語であり，異なり語数で 3484 語の頻度で現れる。最も多く現れるのは「面倒臭い」「照れ臭い」「胡散臭い」「古臭い」などである。1 回だけ現れるのは 175 語で，学習者にその造語方法，パターンを教えると，語彙力の増加に直接繋がると考えられる。

「形容詞＋名詞のコロケーションデータ」は，BCCWJ と JpTenTen から個別に抽出するため，それぞれのコーパスでの高頻度の 500 語の形容詞を見出し語にし，それに，高頻度の長単位の形容詞を追加した。「日本語教育のための形容詞と名詞のコロケーション辞書」は，前述した 62% の形容詞の利用をカバーしている高頻度の基本 25 語の形容詞の記述を目指しているが，本稿では「高い」をモデルにして，作成方法および記述を示す。今後の課題として，同じような方法で，90% の形容詞の利用をカバーしている 127 語の形容詞を見出し語にすることが望ましい。

4.2 コロケーション抽出とコロケーションのシンタクスの検討

3 節で紹介したように，コロケーション抽出は主に中納言（BCCWJ からの抽出）とスケッチエンジン（JpTenTen からの抽出）で行う。ここではコロケーションの頻度か統計的重要度から単純にコロケーションを抽出するだけでなく，シンタクスまで考える必要がある。

---

[12] 予備的分析では，BCCWJ と JpWaC を利用したが，その後新しい超大規模な JpTenTen のウェブコーパスができたので，JpWaC の代わりに JpTenTen のデータを利用する。
[13]「TM」は，作成者の頭文字である。

初期のころのコロケーション研究は，統計から見た単位と単位の組み合わせに焦点を当てていたが，近年の参考研究ではシンタクスおよびシンタクスを超えて考えるまでの方法へのシフトが見られる（例えば，Grefenstette 1992, Cantos & Sánchez 2001, Stefanowitsch & Gries 2003）。McEnery & Hardie（2012）が指摘しているように，コロケーションにおけるパターンおよびシンタクスによく対応しているツールがスケッチエンジンである。このツールの新しいアプローチはシンタクスを考えたパターンの規則化を採用していることで，第 4 世代のコンコーダンスツールであると言える。

形容詞と名詞のコロケーションの抽出に当たり，形容詞連体形と名詞を検索すると，その形容詞が前の文節に対する述語になっているケース（例えば「人口が多い国」）と，形容詞の前に名詞が付く言語表現（例えば「悩み多き年頃」）の場合がある。そこで，連体形の形容詞と名詞の用法を区別するために，パイロットスタディを行った（スルダノヴィッチ 2013）。ランダムに取り出した形容詞と名詞のコロケーションデータのコンコーダンスの例を分析し，単独名詞とそれを修飾する形容詞の組み合わせ以外に，そのコロケーションが文節・文章においてどのような構成になるかを検討した。その分析に基づいて，最もよい抽出結果を出すために以下に示す「DUAL」というルールを作成した。このルールにより，述語の役割を持つ連体形の形容詞を除外することが可能になった。

スケッチエンジンは，ルールの設定によって，どのような単位の組み合わせ，どのようなシンタクスで抽出したいかを指定することができる。以下の形容詞と名詞を取り出すルールは，コロケーションデータを分析しながら調整し，必要に応じて変更を加えた。

```
*DUAL
=modifier_Ai/modifies_N
2: [tag="Ai.*" & word!="ない｜無い" & infl_form="Attr.*"] [tag="Pref"]? 1:[tag="N.*" & tag!="N.num"]
within! [word="が｜の" | tag="N.*"] [tag="Ai.*" & word!="ない｜無い" & infl_form="Attr.*"]
[tag="Pref"]? [tag="N.*" & tag!="N.num"]
```

ここに挙げたのは 2 単位を取り出すためのルールである。2 は，連体形の形容詞（「ない」「無い」を除外）を取り出し，1 は，名詞を取り出す（数詞を除外）。ただし，この形式の前に「が・の・名詞」が無い場合に限る[14]。

このルールによってスケッチエンジンで取り出されるコロケーションを，図 1 の左にある Modifies_N のコロケーションタイプに示す。図 1 は，ワードスケッチ機能で「高い」という形容詞を検索した結果のスクリーンショットである。

さらに，形容詞＋名詞の前後文脈を見せるために，別のルールを作成した。それによって，コロケーションリンクからその前後文脈の高頻度の単位に飛ぶことができる。

[14]「が」「の」「名詞」＋連体形の形容詞＋名詞を扱うルールは別のルールにしたので，このコロケーションデータも抽出できる。

中納言の検索条件は，キーが品詞として名詞，その前方コロケーションが品詞として形容詞で，活用形として連体形，語彙素として対象の形容詞を指定する。他の必要な条件は，テキストとしてダウンロードした上で，フィルターで整理する。

高い jpTenTen11 [SUW] freq = 3842021 (372.2 per million)

modifies_N	585081	-5.8
所	33988	6.17
評価	30911	8.38
物	23027	4.75
レベル	13553	7.08
位置	13548	7.45
金	9636	5.94
事	9148	2.14
方	8805	3.59
技術	8745	6.27
声	7709	5.49
人気	5308	5.7
山	5268	5.83
確率	5001	7.34
場所	4834	5.05
値段	4575	6.64
買い物	4467	6.57
音	4441	5.07

suffix	584795	-0.7
さ	571162	11.17
過ぎ	9959	7.59
認	415	4.52
ちゃん	314	0.72
等	273	1.08
ばい	272	1.67
校	84	1.65
防	75	2.03
源	72	1.58
植え	43	1.26
がる	40	0.2
出	36	0.74
酷	31	0.79
建て	30	0.41
放し	29	0.2
っこい	27	0.38
似	23	0.3

NがのAi	530881	
NがのAi+_物_	26799	-17.1
NがのAi+_事_	22953	-9.6
NがのAi+_人_	13937	-10.8
NがのAi+_為_	11488	-11.8
NがのAi+_方_	10196	-12.4
NがのAi+_所_	9677	-5.6
NがのAi+_作品_	5066	-33.7
NがのAi+_商品_	4566	-47.6
NがのAi+_場合_	3686	-8.1
NがのAi+_サービス_	3463	-71.2
NがのAi+_時_	3382	-4.3
NがのAi+_場所_	3197	-10.5
NがのAi+_訳_	2926	-7.9
NがのAi+_製品_	2830	-59.2
NがのAi+_位置_	2722	-48.9
NがのAi+_地域_	2682	-20.1
NがのAi+_技術_	2559	-65.4

のAi+N	441574	-13.1
物	18612	4.45
人	9041	3.48
所	6253	3.74
方	5641	2.96
作品	4326	4.89
商品	3857	4.68
サービス	3366	4.76
技術	2665	4.61
製品	2544	5.17
情報	2354	3.32
場所	2338	4.05
事	2250	0.12
サイト	1912	3.47
日本	1868	2.33
選手	1775	3.88
評価	1678	4.25
地域	1632	4.24

-N-のAi+N	426845	-14.4
性	75111	8.68
度	41169	6.95
質	25227	9.22
人気	19018	7.61
背	17453	9.15
レベル	15456	7.34
効果	11764	6.61
率	10180	6.75
クオリティー	9272	8.72
価値	9007	7.39
力	7255	5.32
精度	7224	8.32
評価	6075	6.11
頻度	5071	7.89
能力	4293	5.84
リスク	3867	6.77
意識	3708	5.46

図 1　コロケーション抽出の例―「高い」と共起する名詞，接尾辞などの単位―

4.3 2 種のコーパスにおけるコロケーションの比較

BCCWJ と JpTenTen という 2 種のコーパスから抽出したコロケーションを比較してみると，共通点と相違点が認められる。特に，高頻度・中頻度のコロケーションデータには共通点が非常に多く見られる。両方の大規模コーパスから同じ結果が導き出されるということは，データの有意義さおよびコーパスの信頼性を示している。得られたデータのうち，差異が確認できる場合には，それぞれのコーパスが持つ特殊性を確認することができる。BCCWJ は，サブコーパスごとの比較ができるので，差異が現れる場合，ジャンルごとのコロケーションの用法が確認できる。

表 1 に，両方のコーパスにおける「高い＋名詞」のコロケーションを示す。取り出した高頻度の 100 語のうち，特に高頻度のコロケーションが 38 語見られる。100 語のうちで，どちらか片方のコーパスにしか出現していない語は，網掛けでマークされている。「伸び」「比率」「経済」は，BCCWJ の高頻度の 100 語に入っているが，JpTenTen では高頻度 100 語のうちに入っていない。これらは，政府の報告書に現れる用語や経済的用語だと考えられることから，サブコーパスごとのデータを調べたところ，特定目的・白書のサブコーパスに高頻度で現れており，出現に偏りがあることが分かる。一方，「品質」「クオリティー」「奴」は，JpTenTen の高頻度 100 語のうちにのみあることから，ウェブデータ全体に渡って分布していると思われる。「品質」「クオリティー」は品物の質を表す語であるが，これはウェブ上にネットショップの類が多く存在するためと考えられる。「奴」は，ウェブのインフォーマルなスタイルを示している。

表 1　JpTenTen と BCCWJ における「高い＋名詞」の上位 100 語のコロケーションの比較（一部）

JpTenTen			BCCWJ		
Adj+N(3842021)	頻度	相対頻度	Adj+N(41632)	頻度	相対頻度
所	33988	8,85	所	493	11,84
評価	30911	8,05	物	391	9,39
物	23027	5,99	評価	237	5,69
レベル	13553	3,53	水準	200	4,80
位置	13548	3,53	伸び	166	3,99
金	9636	2,51	事	162	3,89
事	9148	2,38	レベル	139	3,34
方	8805	2,29	山	138	3,31
技術	8745	2,28	位置	120	2,88
声	7709	2,01	声	92	2,21
人気	5308	1,38	方	90	2,16
山	5268	1,37	割合	89	2,14
確率	5001	1,30	地位	88	2,11
場所	4834	1,26	比率	84	2,02
値段	4575	1,19	値段	64	1,54
買い物	4467	1,16	技術	63	1,51
音	4441	1,16	場所	59	1,42
効果	4099	1,07	数値	56	1,35
為	4067	1,06	音	48	1,15
信頼	3777	0,98	木	47	1,13
水準	3738	0,97	天井	46	1,10
品質	3680	0,96	値	44	1,06
訳	3196	0,83	価格	44	1,06
建物	2970	0,77	目標	41	0,98
気	2806	0,73	成長	40	0,96
数値	2662	0,69	為	36	0,86
値	2629	0,68	時	35	0,84
価格	2599	0,68	確率	35	0,84
天井	2515	0,65	建物	35	0,84
クオリティー	2509	0,65	経済	35	0,84
精度	2419	0,63	価値	35	0,84
筈	2387	0,62	能力	34	0,82
支持	2354	0,61	教育	33	0,79
ハードル	2300	0,60	訳	33	0,79
目標	2299	0,60	金額	32	0,77
ビル	2292	0,60	関心	32	0,77
次元	2275	0,59	次元	32	0,77
奴	2099	0,55	塀	29	0,70

4.4「形容詞と名詞のコロケーションデータ」の記述

本節では，「形容詞と名詞のコロケーションデータ」の記述および特徴について紹介する。「形容詞と名詞のコロケーションデータ」は，以下 3 種を作成する。

① BCCWJ を利用した形容詞と名詞のコロケーションデータ
② JpTenTen を利用した形容詞と名詞のコロケーションデータ
③ BCCWJ と JpTenTen を利用した形容詞と名詞のコロケーションデータの比較

それぞれのデータでは，高頻度の 500 語の形容詞を見出し語にして，コロケーションデータを示す。取り出すコロケーション語数は，形容詞の頻度，形容詞の用法傾向，コーパスのサイズなどと関連しており，形容詞ごと，またはコーパスごとに違いがある。そのため，最多の場合には，ある形容詞の 100 語のコロケーションのデータおよびその代表的な前後文脈を表示する。一方では，名詞とのコロケーションがない場合がある。これは，形容詞が名詞を修飾する役割を持っていない，およびその役割が非常に少ないという場合である。例えば，「手早い」，「所狭い」などであり，形容詞の 500 語のうち，このような形容詞はあわせて 28 語ある（JpTenTen のデータ）。これらの形容詞には「手早く片付ける」「所狭く並んである」など副詞的な用法が多い。終止形，連体形，連用形は，用法の量的な面では偏りがあり，今後の詳細分析には興味深い課題であり，日本語教育のためにも重要な情報になる。

データ①と②の形式は，JpTenTen から取り出した「高い / 凄い / 悪い / 早い＋名詞」を例にして，それぞれの形容詞の 100 列のうち，最上の 5 列を表 2 と表 3 に示す。表 2 はデータの左側の形容詞の頻度，形容詞＋名詞の組み合わせの頻度，形容詞＋名詞の前文脈のデータの一部を示す。

表 2　形容詞と名詞のコロケーションデータの一部（左側）

Ai頻度	Ai+N頻度	前脈3+2+1	前脈2+1	前脈1	Ai	N
3842021	33988	ま し た 。/の で 、/と 煙 は	の 一番/た 。/ま す 。	、/。/は	**高い**	所
3842021	30911	と し て/世界 的 に/海外 で も	し て/で も/から も	で/も/に	**高い**	評価
3842021	23027	は 非常 に/が 非常 に/、 そんな に	非常 に/そんな に/タ ダ より	は/に/、	**高い**	物
3842021	13553	、 非常 に/は 、 より/は 非常 に	、 より/非常 に/は 、	より/に/、	**高い**	レベル
3842021	13548	ま し た 。/の で 、/を 心臓 より	より も/た 。/は 、	、/より/も	**高い**	位置
…	…	…	…	…	…	…
3808164	57103	。 これ は/だけ で も/いう の は	の は/これ は/本当 に	は/、/に	**凄い**	事
3808164	27036	ま し た 。/・ ・ ・/し て も	本当 に/た 。/に は	、/は/。	**凄い**	人
3808164	17068	ま し た 。/！ 」 と/です が 、	と 、/て 、/が 、	、/は/。	**凄い**	勢い
3808164	16811	いう の は/は 本当 に/って の は	の は/と は/力 は	は/、/に	**凄い**	物
3808164	6821	ま し た 。/て て 、/・ ・ ・	た 。/と か/て 、	、/。/も	**凄い**	格好
…	…	…	…	…	…	…
2931564	91894	いい こ と と/は 何 も/良い こ と も	の は/何 も/何 か	、/は/も	**悪い**	事
2931564	17591	良い ところ も/いい ところ も/良い ところ 、	ところ も/ところ 、/ところ と	、/も/。	**悪い**	所
2931564	16038	意味 で も/で も 、/ま し た 。	で も/も 、/は 、	も/、/。	**悪い**	意味
2931564	14543	良い もの と/、 体 に/良い もの も	体 に/何 か/もの と	に/、/は	**悪い**	物
2931564	12627	良い 奴 、/いい 奴 、/は そんな に	本当 に/奴 、/そんな に	、/は/に	**悪い**	奴
…	…	…	…	…	…	…
2524026	24960	経つ の は/ま し た 。/たつ の は	の は/た 。/ま す 。	。/は/、	**早い**	物
2524026	17160	の で 、/、 比較 的/、 できる だけ	比較 的/は 、/、 かなり	、/は/かなり	**早い**	段階
2524026	16108	いつ も より/の で 、/ま し た 。	に は/た 。/もう 少し	、/は/。	**早い**	時間
2524026	12239	、 できる だけ/の できる だけ/、 比較 的	できる だけ/比較 的/は 、	、/は/かなり	**早い**	時期
2524026	11785	の で 、/、 できる だけ/ま し た 。	で 、/は 、/できる だけ	、/は/。	**早い**	内
…	…	…	…	…	…	…

表 3 は，形容詞＋名詞の後文脈のデータ，形容詞の出現形の一部を示す。前後文脈は，コーパス中に最も高い頻度で現れている 3 例ずつを表示する。表 2 の「前脈 1」は，コロケーションの前文脈の代表的な 1 単位（「高い位置」を例にとると　、/ より / も），「前脈 2+1」は，代表的な 2 単位（より も / た 。/ は 、），「前脈 3+2+1」は代表的な 3 単位（まし た 。/ の で 。/ を 心臓 より）。前後文脈は同じように示される。これで，「高い位置」は文章・文節の初めによく現れ，「を心臓より高い位置」のパターンによく見られることが分かるが，もっと細かい情報はさらにコーパスで調べる必要がある。普段 1 単位として使われているが，形態素解析で短単位に分けた単位も文脈で把握できる（「早い時間＋帯，高い技術＋力」）。このデータは，現時点で 500 語の形容詞とその名詞とのコロケーションデータ（23247 語）を JpTenTen から取り出した。

データ③は，「高い」を例にして，それぞれのコーパスから抽出したコロケーションデータを揃えて，頻度，計算したコーパスごとの相対頻度とその差を比較できるような表示にして，表 4 に示した。表 4 では，片方のコーパスにしか出てこない語は特別にマークした上で，最も高頻度の 100 語のコロケーションのうち，2 つのコーパスにおける最初の 50 語を比較する。（比較のメリットについては 4.3 節を参照されたい。）

以上で示した，日本語の形容詞＋名詞のコロケーションデータは，従来作成されていなかったものである。今後，日本語研究，日本語教育，辞書学などの分野において幅広く活用できるものと考えられる。例えば，日本語研究においては，500 語の形容詞と名詞と前後文脈のデータから形容詞の様々な分法的･意味的振る舞い，形容詞･名詞の意味的特徴により共起傾向推移のパターン，前後文脈に現れる副詞，比較級の傾向などについてさらに分類できる。日本語教育・辞書学においては，次節の「日本語教育のための形容詞と名詞のコロケーション辞書」で示すように，

表 3　形容詞と名詞のコロケーションデータの一部（右側）

Ai	N	後脈1	後脈1+2	後脈1+2+3	単語
高い	所	に/から/で	に ある/が 好き/に いる	が 好き な/に ある の/から 落ち て	高い/高き/たかい
高い	評価	を/が/と	を 得/を 受け/を 受ける	を 得 て/を 受け て/を 得 た	高い/たかい/高き
高い	物	で/を/は	で は/の 、/で も	で は ない/に なっ て/が あり ます	高い/たかい/高き
高い	レベル	で/の/に	で の/に ある/で 実現	で 実現 し/で 安定 し/に ある 。	高い/たかい/高～い
高い	位置	に/で/から	に ある/で ボール/から の	で ボール を/に ある の/に あり 、	高い/たかい/高き
…	…	…	…	…	…
凄い	事	に/だ/です	に なっ/だ と/に なる	に なっ て/に なっ てる/だ と 思う	すごい/凄い/スゴイ
凄い	人	だ/が/です	だっ た/で し た/な ん	で し た 。/な ん だ/だっ た 。	すごい/凄い/スゴイ
凄い	勢い	で/です/の	で 食べ/で 、/で 走っ	で 食べ て/で 走っ て/で 売れ て	すごい/凄い/すっごい
凄い	物	が/を/だ	が ある/が あり/です 。	が あり ます/が ある 。/で し た 。	すごい/凄い/スゴイ
凄い	格好	いい/よかっ/イイ	よかっ た/良かっ た/よく て	よかっ た 。/よかっ た です/よかっ た ！	すごい/すっごい/凄い
…	…	…	…	…	…
悪い	事	を/し/は	を し/で は/し た	で は ない/を し た/を し て	悪い/わるい/ワルイ
悪い	所	は/が/を	が ある/は ない/も ある	で も ある/が あれ ば/が あっ た	悪い/わるい/ワルイ
悪い	意味	で/じゃ/の	で も/で は/で の	で は なく/で も 、/で は ない	悪い/わるい/惡い
悪い	物	で/を/は	で は/で も/じゃ ない	で は ない/で は あり/で は なかっ	悪い/わるい/ワルイ
悪い	奴	ら/は/が	じゃ ない/で は/は い	で は ない/は い ない/、 変 な	悪い/わるい/ワルイ
…	…	…	…	…	…
早い	物	で/です/勝ち	で 、/です 。/で もう	です ね 。/で 、 もう/で 今年 も	早い/速い/はやい
早い	段階	で/から/に	で 、/で の/から 、	で 自分 の/で 出 て/で 気づい て	早い/速い/はやい
早い	時間	に/から/帯	帯 に/だっ た/帯 で	だっ た の/な の で/な の に	早い/速い/はやい
早い	時期	に/から/の	に 、/から 、/から の	に やっ て/です が 、/な の で	早い/はやい/速い
早い	内	に/から/の	に 、/に 手/から 、	に 手 を/に やっ て/に 一 度	早い/はやい/速い
…	…	…	…	…	…

表 4　形容詞と名詞のコロケーションデータの比較の一部（「高い」の例）

	2 種のコーパス	JpTenTen 相対頻度	BCCWJ 相対頻度	合計	差
1	所	8.85	11.84	20.69	3.00
2	物	5.99	9.39	15.39	3.40
3	評価	8.05	5.69	13.74	-2.35
4	レベル	3.53	3.34	6.87	-0.19
5	位置	3.53	2.88	6.41	-0.64
6	事	2.38	3.89	6.27	1.51
7	水準	0.97	4.80	5.78	3.83
8	山	1.37	3.31	4.69	1.94
9	方	2.29	2.16	4.45	-0.13
10	声	2.01	2.21	4.22	0.20
11	伸び	0.00	3.99	3.99	3.99
12	技術	2.28	1.51	3.79	-0.76
13	金	2.51	0.67	3.18	-1.84
14	値段	1.19	1.54	2.73	0.35
15	場所	1.26	1.42	2.68	0.16
16	地位	0.41	2.11	2.53	1.70
17	割合	0.29	2.14	2.43	1.85
18	音	1.16	1.15	2.31	0.00
19	確率	1.30	0.84	2.14	-0.46
20	数値	0.69	1.35	2.04	0.65
21	比率	0.00	2.02	2.02	2.02
22	為	1.06	0.86	1.92	-0.19
23	人気	1.38	0.50	1.89	-0.88
24	天井	0.65	1.10	1.76	0.45
25	値	0.68	1.06	1.74	0.37

	2 種のコーパス	JpTenTen 相対頻度	BCCWJ 相対頻度	合計	差
26	買い物	1.16	0.58	1.74	-0.59
27	価格	0.68	1.06	1.73	0.38
28	訳	0.83	0.79	1.62	-0.04
29	建物	0.77	0.84	1.61	0.07
30	目標	0.60	0.98	1.58	0.39
31	木	0.42	1.13	1.55	0.71
32	効果	1.07	0.48	1.55	-0.59
33	信頼	0.98	0.41	1.39	-0.57
34	次元	0.59	0.77	1.36	0.18
35	能力	0.51	0.82	1.33	0.30
36	ビル	0.60	0.70	1.29	0.10
37	金額	0.51	0.77	1.28	0.26
38	成長	0.25	0.96	1.21	0.71
39	関心	0.43	0.77	1.20	0.33
40	時	0.36	0.84	1.20	0.48
41	気	0.73	0.46	1.19	-0.27
42	場合	0.51	0.65	1.15	0.14
43	筈	0.62	0.53	1.15	-0.09
44	価値	0.28	0.84	1.12	0.56
45	支持	0.61	0.48	1.09	-0.13
46	温度	0.42	0.65	1.07	0.23
47	給料	0.55	0.50	1.05	-0.04
48	周波	0.40	0.62	1.03	0.22
49	人	0.47	0.55	1.02	0.08
50	ハードル	0.60	0.41	1.01	-0.19

コーパスから取り出した形容詞と名詞の基礎データを基にして，さらに様々な要因を考慮に入れて，コーパス分析の結果を取り入れながら，学習者用の教材作成・辞書編纂に応用することが可能である。

4.5「日本語教育のための形容詞と名詞のコロケーション辞書」の作成と記述

本節では，前節で述べてきた「形容詞と名詞のコロケーションデータ」を基にした「日本語教育のための形容詞と名詞のコロケーション辞書」の作成方法と記述を紹介する。モデルとして「高い」という形容詞を取り上げ，その記述を図 2 に示す。

タカイ
高い

4級	トコロ 所	ヤマ 山	**コエ** **声**	**カネ** **金**	カイモノ 買い物	タテモノ 建物	キ 木	**セガ～ヒト** **(背が～)人**			
3級	**セガ～カタ** **(背が～)方**	ギジュツ(カ) 技術(力)	ネダン 値段	バショ 場所	ワリアイ 割合	**オト** **音**	ビル ビル	**カベ** **壁**	**キョウイク** **教育**		
2級	ヒョウカ 評価	レベル レベル	イチ 位置	スイジュン 水準	チイ 地位	カクリツ 確率	スウチ 数値	ニンキ 人気	テンジョウ 天井	カカク 価格	モクヒョウ 目標
	コウカ 効果	シンライ(セイ) 信頼(性)	ノウリョク 能力	キンガク 金額	セイチョウ(リツ) 成長(率)	カンシン 関心	カチ 価値	オンド 温度	キュウリョウ 給料	ヘイ 塀	
1級	ヒリツ 比率	アタイ 値	シジ（率） 支持(率)	ヒンシツ 品質							
級外	ノビ(リツ) 伸び(率)	(ヨリ～)ジゲン (より～)次元	シュウハ(スウ) 周波(数)	ハードル ハードル	セイド 精度	キンリ 金利					

high, tall, expensive

4th	high place	high mountain	**high-pitched** **voice**	**lots of** **money**	expensive purchase	tall building	high tree		**tall** **person, man**		
3th	**tall** **person**	high technology, skills	high, expensive price	high location	high percentage	**high-pitched** **sound**	high building	high **wall**	high-quality **education**		
2nd	high evaluation	high level	high position	high level	high status	high probability	high figures	high popularity	high ceiling	high, expensive price	high goal
	high effect	high reliability	high ability	high, great price	high growth	high, great interest	high value	high temperature	high salary	high wall, fence	
1st	high ratio	high value	high support(-ratio)	high quality							
zero	high growth	high(er) dimension	high frequencies	high bar, hurdle	high accuracy	high interest rate					

図 2 「高い」を見出し語にした「日本語教育のための形容詞と名詞のコロケーション辞書」の記述例

1）コロケーションリストの整理

2 種のコーパスにおける相対頻度で比較して選んだ高頻度の 50 語のコロケーションは「形容詞と名詞のコロケーションデータ」から得られるが，日本語教育への応用の面からコロケーションリストを観察して，整理する必要がある。例えば，「高い」と共起する名詞のうち，機能語として使われている単語（「物」「事」「為」「訳」「筈」など）は，別のデータとして扱い，辞書の対象になるコロケーションリストから排除した。一方，「方」「所」のような一般名詞としても機能語としても使われている単語はリストに残した。さらに，コロケーションの前後文脈を観察した結果，形態素解析では短単位に分けたが，他の単位との組み合わせでよく現れている語の場合には，それを括弧によって示した。例えば，「(背が）高い人」，「技術（力)」，「信頼（性)」などの「高い」のコロケーション例が見られる。なお，各コロケーションには，学習者が読み方を確認できるように，振り仮名を付ける。

2）コロケーションを構成する要素の難易度の判定

日本語教育への応用を考慮に入れて，学習者にとって重要な情報である難易度を判定し，辞書に加えた。コロケーションの難易度をそのまま判定できるリソースはないので，第一のステップとしてコロケーションを構成するそれぞれの単位の難易度を旧 JLPT の 4 段階のレベルに分ける。JLPT 語彙リストにない項目は，級外にする。それによって，辞書を利用する学習者が自分の学

習レベルを判断し，既に習得している単語のうち，どの単語を組み合わせることができるかを簡単に知ることができるようになっている。例えば,「高い」の場合には, 4 級になっている単語は，位置（「所」），自然（「山，木」），人と関連しているもの（「人，声」），人が作った建築（「建物」），販売と関連しているもの（「金，買い物」）などである。なお,「高い買い物」「高い金」のように，コロケーションを構成する各単語はやさしいものの，組み合わせとしては初級では扱いにくいコロケーションがある。このようなコロケーションの難易度については個別に注記する。さらに，抽象的な概念を表している 2 級の単語が多いということに気付くことができる。

3）予想しにくい，または予想しやすいコロケーションの検討およびノートの執筆

コロケーションを予想できるかどうかは，言語によって異なるが，ある言語においてのみ特殊な振る舞いをするコロケーションは，多くの言語から見ると，予想し難いと考えられる。本研究で記述するコロケーションは，英語話者の学習者を前提にしてコロケーションの予想を決定している。その基準は，日本語のコロケーションを直訳した時に英語のコロケーションとして成立するかどうか，すなわち，形式，意味，コロケーションの単位の組み合わせに違いがあるかどうかという点である。辞書では，予想しにくいコロケーションについては，網掛けでマークする。

図 2 に示したように，学習者が分からない単語の意味をすぐに理解できるように，英語訳を付ける。英語訳を観察すると，予想できるコロケーションと予想できないコロケーションの違いを見て取ることができる。図 2 では，予想しにくいのは，1）形式が違う（背が高い人 “a tall person”,「高い人」だけでは伝わらない), 2）母語でその組み合わせにならないコロケーション（高いお金 “lots of money”），3）母語では，その組み合わせがあるが，意味の違いがある（高い教育 “a high-quality education/high education”），という 3 つのパターンが見られる。

予想しにくいコロケーションについては，特に注意が必要なため，コーパスでパターンを分析し, 他のコロケーションとの振る舞いを検討し, 辞書にその用法についてさらに記述する（図 3）。例えば，学習者が混乱しやすい「高い声」と「大きい声」がどのように違うか,「高いお金」「たくさんのお金」「多くのお金」にどのような用法があるか, などである。それらの用法を示すため，コーパスから例を取り出す，または，誤用に関する注意を書き加える。例えば,「大きい声で言ってください。（注：このよく使われる表現は「高い」が使えない。）」。1 つの例としてコーパスから取り出した「多くのお金を」と「高いお金を」の違いを図 4 に示す。意味・用法と関連する動詞との共起パターンの違いがあるので，ノートに記述する。

高い声
「**高い声**」と「**大きい声**」は、意味・用法が違う。
「高い声」＋で＋「歌う、鳴く、叫ぶ」(with a high-pitched voice)
「大きい声」＋で＋「言う、話す、歌う、叫ぶ、挨拶する」(with a loud voice, loudly)
● カラオケで**高い声** がでなくなって辞めた 。| 若いメイド達が悲鳴の様な**高い声**を上げた。| 以前 出ていた**高い声**が出なくなったらどう思いますか ？
● **大きい声で**言ってください。(注：このよく使われる表現は「高い」で使えない。) | **大きい声で**は言えないような話も多い。

背が高い人
「高い＋人」「高い＋方」は、「**背が高い人**」「**背が高い方**」(a tall person) の表現でよく使われる。
それ以外に、以下のような用法がある。(「高い」の連体修飾の用法については次のページをご覧ください)
● 社会的地位の**高い人** | 自分よりレベルの**高い人**
● (結婚・安全...)～に対する意識の**高い方** | コミュニケーション能力の**高い方**

高いお金
「**高いお金**」(lots of money)は、「払う、出す、掛ける」という動詞とよく使われる。
● **高いお金**を払って、いい大学に行きたい。
一方、「**たくさんお金を**」＋「使う、稼ぐ、掛ける、払う、持つ」、および「**多くのお金を**」＋「稼ぐ、費やす、使う、得る、儲ける、作る、持つ、節約する」の利用がある。
● **たくさんお金**を使って、経済を回さないといけないんです。| 短期間で出来るだけ**多くのお金**を稼ぎたいと思います 。

高い壁
「**高い壁**」(a high wall, a high-walled,high barriers)は、高い建物の部分の意味もたまにあるが、抽象的な意味のほうが多い。
● 円形の舞台を**高い壁**が取り囲んでいる。
● 三位と四位の間にはとても**高い壁**が存在する。| 大学への進学は経済的に困難と分かり、人生で初めて「**高い壁**」を感じる。

ジャンル情報
「**高い伸び**」「**高い伸び率**」は、経済的用語としてよく使われる。

図 3　予想しにくいコロケーションの記述例（ノート）

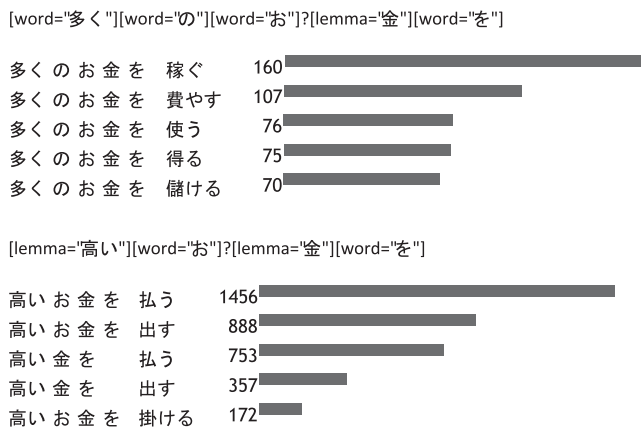

図 4　スケッチエンジンで取り出したパターン―「多くのお金を」と「高いお金を」―

4）コロケーションの語彙マップの作成

日本語教育において形容詞および名詞の習得に役に立てるため，また意味的・認知的に見出し語の用法が把握できるようにするため，各形容詞と共起する複数の名詞のデータを意味セットに分けて，語彙マップのモデルを作成した。3 節で述べたように，認知的・意味的に検討し，記述された形容詞についての参考文献（今井 2011，西尾 1972，八亀 2008，国立国語研究所編 2004）を参考にしつつ，日本語教育の立場および認知言語学の立場から，語彙マップのグルーピングおよび順番を判断する。

図 5 に示すように，ぞれぞれのコロケーションは，意味によってセットになっているため，「高い」がどのような多義性を持っているかが分かる。大まかに，その用法は，位置関係，量的関係，優劣の 3 つに分かれている。後者の 2 つは，位置関係からの比喩による広がりであると思われる。これらの用法の意味領域情報は教育にも応用できると思われる。位置関係の 1 つ目のグループは，具体的なもので，自然に見られる山・樹木か人が作成した建物・建物の一部である。1 つの意味セットの中では，高頻度のデータから当該形容詞に近い順にコロケーションが並んでいる（山）。または，具体的な意味のほうが中心に近づいており（山，木，ビル，建物），抽象的な意味・比喩的な意味（壁，ハードル）は，遠くなっている。続いて，人および人と関連している声・音のグループである。次は，位置と直接関連している所，場所，位置のグループおよびそこから抽象的な意味になってきた社会・組織の中での地位である。量的関係は，数，程度，率を示すグループ（確率，割合など）とお金と関連しているグループである（金，金額など）。優劣は，質・評判・価値のグループ（評価，レベルなど），能力のグループ（能力，技術力），意識・理想（目標，関心など）が優れているグループからなっている。

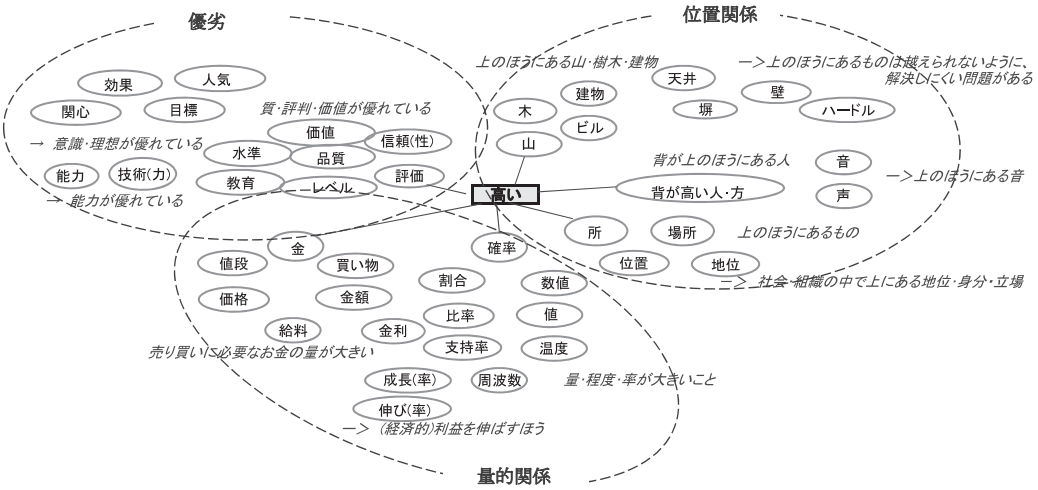

図 5　名詞のコロケーションを記述した「高い」の語彙マップ（試案）

語彙マップにしたコロケーションとしての名詞から，形容詞の用法がさらに分かり，また，セットになっている名詞の学習にも役立てることができる。得られたコロケーションデータを，視覚的にも意味的にも明白で魅力的な内容として語彙マップで体系的に表示することによって，学習者に対して使いやすさと学習しやすさを提供することができる。語彙マップの有意義性を明らかにしている研究としては，Morin & Goebel（2001）が挙げられる。

5）形容詞の用法についての検討およびジャンル情報，追加情報の執筆

その後の追加情報は，対象の形容詞の文法・意味的特徴などである。例えば，「高い」は連体

修飾の利用は 50% ぐらいで，その最もよく現れる用法は「（信頼・安全・可能）性，（完成・難易・自由）度，質，人気，背など＋の＋高い＋名詞」である。

ジャンルごとの差異の情報は，2 種のコーパスおよびサブコーパスのデータを確認した結果，記入する。例えば，「高い伸び」「高い伸び率」は，BCCWJ の白書のデータに偏りがあり，経済的用語としてよく使われる表現である。

5. 理論・応用におけるデータの意義

本稿で提唱したコロケーションデータが完成すれば，それを基にして，言語学・第二言語教育の分野における様々な理論的・応用的研究を発展させられる可能性がある。本節では，いくつかの例を取り上げ，今後の可能性について論じる。

5.1 名詞と形容詞のコロケーションとシンタクス

コロケーションを取り出すための方法論は，統計的アプローチおよびコンコーダンサーにおける用法・意味・パターンの観察方法が多く用いられてきた。これに加えて，コロケーションをシンタクスの面から観察すると，さらに充実したコロケーション用法の記述ができる。従来の研究では，文中の形容詞の機能について多くの指摘があり，主として，叙述用法，連体用法，連用用法の 3 つに分けて，どの機能が中心的であるかについて議論されてきた（西尾 1972，宮島 1993，八亀 2008 など）。コーパスに基づいた研究としては，形容詞の用法分布・語義分布についての考察（姜 2012），連体形か連用形でしか利用されない形容詞（小川他 2008），形容詞述語のタイプと述語になりやすい語となりにくい語（前川 2012）などが挙げられる。

このプロジェクトの研究の一部として，「形容詞（連体形）＋名詞」のコロケーションを対象にして，単独名詞とそれを修飾する形容詞の組み合わせ以外に，どのような構成がありうるかについて，高・中・低頻度の形容詞（各 3 語）を検討した（スルダノヴィッチ 2013）。その結果，以下のものがあることが分かった。a）形容詞単独で名詞を修飾する（「寒い季節」），b）連体修飾節の述語として名詞を修飾する（「雨の多い国」），c）所有の「の」＋形容詞で名詞を修飾する（「男性の野太い声」），d）複合形容詞および連体修飾の「が」の省略で名詞を修飾する（「香り高いコーヒー」）。また，それぞれの形容詞に振る舞いの違いがあることが明らかになった。例えば，「多い」は述語の機能の出現は非常に高く，9 割を超えている。「高い」は，連体形の形容詞が連体修飾節の述語であるケースが全体のおよそ半数を占めており，形容詞の前に名詞が付く言語表現が 5 例あった（「香り高いコーヒー」）。一方，「青い」は，連体形の形容詞が連体修飾節の述語になるケースがほとんどなく，所有の「の」の用法が多く見られる（「日本の青い空」）。

4.2 節で述べたようにコロケーションを抽出するためのルールを改良した。すると，「高い＋コーヒー」，「高い＋サービス」のような，確率的な方法で取り出すことはできるが，不十分だと考えられるコロケーションは，さらに正確に取り出せるようになる（「香り高いコーヒー」，「質の高いサービス」）。今後も，コロケーション分析に統語的アプローチを取り入れつつ，実証的および確率的に，語およびその組み合わせの振る舞いを検討し，記述することが望ましい。

5.2 難易度・親密度・連想データ

3.4 節でコロケーションを構成する各単語の難易度を判定するリソースについて述べた。コロケーション全体の難易度を示すリソースは現時点で存在しないため，今後の有意義な研究課題として考えられる。

あるコロケーションの各単位が 4 級であっても，そのコロケーションは 4 級になるわけではなく，初級者に教えるコロケーションではない可能性がある。例えば，4.5 節で取り上げた「高い＋買い物」は，それぞれの語は初級であるが，組み合わせとして必ずしも初級であるとは限らない。数名の日本人教師の直感によると「高い買い物」は初級ではないという意見があった。このような別々の単語とその組み合わせに関する違いを，それぞれの単語のレベルを見ながら検討する必要がある。必要に応じて，日本語母語話者に対する調査を行うことも効果的であろう。モデルとしての実験を行った上で，『日本語の語彙特性』のような新しいデータベースを作成していくことが課題として考えられる。『日本語の語彙特性』は，様々な語彙データを搭載しているが，そのデータの対象は単語レベルにとどまっているため，今後コロケーションを用いたデータを作成することには大きな意義がある。

先行研究では，心像性（imageability，単語から喚起される感覚イメージの思い浮かべやすさを表す主観的特性）は，一般に具像性（concreteness，物や材料，人などを具体的に示す程度）との相関が高く，出現頻度との相関は低く，親密度（familiarity，単語のなじみの程度を主観的に評価した評定値）との関係は研究ごとに違いがあることが示されてきた（佐久間他 2005: 29–37）。さらに，日本語教育においても，日本語語彙リストにおける難易度レベルと親密度の関連についての研究が行われ，親密度の高い語彙はより基礎的な語彙の中に多いことが示された（松田他 2010）。特に，以下の点を検討することが期待されており，その場合にも本研究で作成しているデータがコロケーションの基本リソースとして利用できる。

1） コロケーションを構成する各単語の難易度から，どの程度そのコロケーションの難易度が予想できるか
2） どのような要因によって個別の単語と複合語単位の間で難易度の変化が起こるのか
3） コロケーションを構成する各単位の親密度・心像性・具像性からどの程度そのコロケーションの親密度・心像性・具像性が予想できるか
4） 各単位とコロケーションの難易度・頻度・親密度・心像性・具像性の関係

なお，コロケーションと連想（association）の関連については，Joyce & Srdanović（2008）で，発表されている。これを形容詞と名詞を対象にして検討するのも今後の課題である。

5.3 予想しにくいあるいは予想しやすいコロケーション

予想しにくいコロケーションと予想しやすいものについては，既に 2.2 節と 4.5 節 3）で述べた。

本研究では，英語話者の学習者を前提にしてコロケーションの予想を決定している。今後は，諸言語のコロケーションデータベースの作成を考えているが，その際，予想しにくいコロケーショ

ンを多言語の観点から検討し，言語と言語の間における共通点と差異を把握し，対照的な研究を行うことが望ましい。以下のものは，特に興味深い課題となる。

1） コロケーションの現象における普遍性
2） 類似した言語におけるコロケーションのほうが，他の言語と比べると，予想しやすいかどうか
3） 離れた言語間においてコロケーションの構成には，違いがあるかどうか

さらに，形容詞と名詞のコロケーションデータは，言語類型論の面から見た形容詞の研究への第一歩になりうる[15]。世界の諸言語の中での，日本語の形容詞，および名詞句における名詞と形容詞の組み合わせの位置づけを，対照言語学的な観点からさらに検討できる。

6. まとめ

本稿では，言語理論から見たコロケーション，およびその第二言語学習における重要性について述べ，「形容詞＋名詞」のコロケーションを対象にした記述的研究の進行中の成果を紹介した。また，大規模な均衡コーパスである「現代日本語書き言葉均衡コーパス」（BCCWJ）と超大規模な現代日本語ウェブコーパス「JpTenTen」を用いた，「形容詞＋名詞」の 2 種のコロケーションリソースの作成方法について述べた。第一のリソースは，「形容詞と名詞のコロケーションデータ」であり，シンタクスを考えた抽出方法および作成した 500 語の形容詞のコロケーションデータ（23247 語の名詞）について述べた。これは日本語学･日本語教育などの分野において基本データとしての応用が期待される。得られたコロケーションデータから「高い」をモデルにして，第二のリソース「日本語教育のための形容詞と名詞のコロケーション辞書」の作成方法を紹介した。その特徴は，2 種のコーパスから取り出したコロケーションに難易度，予想しにくさ，コロケーションの用法，ジャンル情報，意味を考慮に入れた語彙マップなどの記述を加えたことである。

さらに，本稿で取り上げたことについて以下の通りまとめる。

- 現代語のバランスが取れたデータ，または複数のコーパスの利用に基づいた日本語コロケーションデータ・日本語コロケーション学習辞典が必要であること。
- 学習者によるコロケーションの産出・理解は，母語のコロケーション知識の影響を受け，不自然な表現になりやすい。それは，特に予想しにくいコロケーションの場合に起こるため，学習者にとって予想しにくいコロケーションを強調する必要があること。
- 大規模コーパスを 2 種利用し，様々なコロケーションを比較した結果，高頻度・中頻度のコロケーションデータには共通点が非常に多くあったことから，両方のコーパスが実証的データとして利用されることが有益であるという点を確認できたこと。
- 一方，それぞれのコーパスに特徴があり，ジャンルごとの用法を示すサブコーパスや 2

[15] 品詞としての形容詞が存在しない言語もある。例えば，動詞か名詞がその機能を持つ言語として Eastern Ojibwa が挙げられる。

種のコーパスの比較にはメリットがある。特に，共通ではない単語の組み合わせをさらに考察することで，各データセットの差異および特殊なコロケーションの語法についての情報が得られること。

- コロケーションを構成する各単語を日本語能力レベルおよび語彙頻度によって初級から上級までのランク付けを行うことで，学習者のレベルに応じたコロケーションの観察ができ，導入順番にも参考になるデータであること。
- 形容詞と組み合わされる名詞を語彙マップで示すことで，形容詞の意味が及ぶ諸領域が解明され，それを名詞･形容詞およびその組み合わせの学習に応用する材料にできること。
- コロケーションから形容詞と名詞を考えると，形容詞のシンタクス的環境の重要性が明らかになった。形容詞が単独で名詞に掛かる場合と連体修飾節の述語として掛かる場合の違い，さらには，名詞とのコロケーションがないか少ない場合の形容詞の副詞的・述語的な用法の違いの重要性が明らかになる。ここから導かれる教訓がもう 1 つある。すなわちコロケーションは単純に語と語の組み合わせだけで扱うのでは不十分だということである。

本稿の主な意義は，従来存在しなかった，日本語学と日本語教育に利用できるコロケーションデータの作成およびその方法論を提案した点にある。作成しているリソースは，客観的なデータに基づいた形容詞と名詞の組み合わせの記述であり，今後の研究のために様々な理論的，応用的な可能性を持っている。コロケーションとシンタクス，予想しにくいコロケーションの現象，連想とコロケーションの関係，コロケーションの難易度・親密度・心像性・具像性の関連，言語類型論から見た形容詞などの研究課題において，作成されたリソースは，基礎的な研究資源として利用できると考えられる。

さらに，今後の発展として，諸言語と日本語のコロケーションデータベースの作成が挙げられる。言語間の共通点と差異，多言語から見た予想しにくいコロケーションの検討を，次の目標にしていきたい。また，応用的な側面として，得られたデータを，既存の日本語教科書におけるコロケーション関係と比較し，現在の日本語教育におけるコロケーションの扱いを検討することが考えられる。なお，誤用情報を付与した多言語の学習者コーパスの構築と共に，コロケーションに関する学習者の誤用の検討を取り入れることが考えられる。

参照文献

秋元美晴・有賀千佳子（1996）『ペアで覚えるいろいろなことば』東京：武蔵野書院.

Bazell, C.E., J.C. Catford, M.A.K. Halliday and R.H. Robins (eds.) (1966) *In memory of F.R. Firth*. London: Longman.

Cantos-Gomez, Pascual and Aquilino Sánchez (2001) Lexical constellations: What collocates fail to tell. *International Journal of Corpus Linguistics* 6(2): 199–228.

Firth, John (1951) *Papers in linguistics*. Oxford: Oxford University Press.

深田淳（2007）「日本語用例・コロケーション情報抽出システム『茶漉』」『日本語科学』22: 161–172.

Grefenstette, John (1992) Use of syntactic context to produce term association lists for text retrieval. *Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval*, 89–97. New York: ACM.

Halliday, M.A.K. (1966) Lexis as a linguistic level. In: C.E. Bazell et al. (eds.), 148–162.

橋本三奈子・青山文啓（1992）「形容詞の三つの用法：終止，連体，連用」『計量国語学』18(5): 201–214.
橋本和佳（2007）「名詞とそれを修飾する形容詞の関係」『日本語学』26(12): 38–46.
姫野昌子（2004）『日本語表現活用辞典』東京：研究社.
姫野昌子（監）柏崎雅世・藤村知子・鈴木智美（編）（2012）『研究社日本語コロケーション辞典』東京：研究社.
Hoey, Michael (2005) *Lexical priming: A new theory of words and language*. London: Routledge.
曹紅荃（2012）「形容詞と名詞の共起表現から見る学習者言語」仁科喜久子（監）鎌田美千子他（編），47–62.
曹紅荃・仁科喜久子（2006）「中国人学習者の作文誤用例から見る共起表現の習得及び教育への提言―名詞と形容詞及び形容動詞の共起表現について―」『日本語教育』130: 70–79. 日本語教育学会.
堀正広（編）（2012）『これからのコロケーション研究』東京：ひつじ書房.
Hornby, Albert Sydney (1954) *A guide to patterns and usage in English*. Oxford: Oxford University Press.
Hunston, Susan (2002) *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
家田章子（2003）「「コノ～ノニ」と形容詞のコロケーション―コーパスによる共起表現の考察―」『言葉と文化（Issues in Language Culture）』4: 213–226. 名古屋大学.
今井新悟（2011）『日本語多義語学習辞典　形容詞・副詞編』東京：アルク.
James, Carl (1998) *Errors in language learning and use*. London: Longman.
姜　紅（JIANG Hong）（2012）「コーパスに基づく多義語「甘い」の意味再分類及び語義分布調査」『第 1 回コーパス日本語学ワークショップ予稿集』59–68. 国立国語研究所.
Joyce, Terry and Irena Srdanović (2008) Comparing lexical relationships observed within Japanese collocation data and Japanese word association norms. *Cognitive Aspects of the Lexicon,Workshop at the 22nd International Conference on Computational Linguistics*.
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」『情報処理学会自然言語処理研究会』171(12): 67–73.
Kilgarriff, Adam & Michael Rundell (2002) Lexical profiling software and its lexicographic applications—A case study. *EURALEX 2002 Proceedings*, 807–818.
Kilgarriff, Adam, Pavel Rychly, Pavel Smrž and David Tugwell (2004) The Sketch Engine. *Proceedings of EURALEX*, 105–116. France: Université de Bretagne.
金田一秀穂（2006）『知っておきたい日本語コロケーション辞典』東京：学習研究社.
Kjellmer, Goran (1991) A mint of phrases. In: Geoffrey Williams & Sandra Vessier (eds.) *English corpus linguistics*, 111–127. London: Longman.
国立国語研究所（編）（2004）『分類語彙表　増補改訂版』東京：大日本図書.
国際交流基金・（財）日本国際教育協会（2002）『日本語能力試験出題基準（改訂版）』東京：凡人社.
前川喜久雄（2008）「KONONOHA『現代日本語書き言葉均衡コーパス』の開発」『日本語の研究』4 (1): 82–95.
前川喜久雄（2012）「「形容詞＋です」述語の生起要因についての準備的考察」『第 1 回コーパス日本語学ワークショップ予稿集』211-220. 国立国語研究所.
丸山岳彦・山崎誠・柏野和佳子・佐野大樹・秋元祐哉・稲益佐知子・田中弥生・大矢内夢子（2011）『『現代日本語書き言葉均衡コーパス』におけるサンプリングの原理と運用』（国立国語研究所内部報告書 LR-CCG-10-01）.
松田真希子・児玉茂昭・竹元勇太・石坂達也・森篤嗣・川村よし子・山本和英（2010）「コーパスの異なりと単語親密度を活用した日本語共通基礎語彙の抽出」『言語処理学会第 16 回年次大会予稿集』579–582.
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）
McCarthy, Michael and Ronald Carter (2006) *This that and the other: Multi-word clusters in spoken English as visible patterns of interaction*. Oxford: Oxford University Press.
McEnery, Tony and Andrew Hardie (2012) *Corpus linguistics: Method, theory and practice*. Cambridge textbooks in linguistics. Cambridge: Cambridge University Press.
McIntosh, Angus (1966) Patterns and ranges. In: Angus McIntosh & M.A.K. Halliday (eds.) *Patterns of language: Papers in general descriptive and applied linguistics*, 183–199. London: Longman.
宮島達夫（1993）「形容詞の語法と用法」『計量国語学』19(2): 94–104.
茂木俊伸（2012）「文法的視点からみた外来語：外来語の品詞性とコロケーション」陣内正敬・相澤正夫・田中牧郎（編）『外来語研究の新展開』46–61. 東京：おうふう.
Morin, Regina and Joseph Goebel Jr. (2001) Basic vocabulary instruction: Teaching strategies or teaching words? *Foreign*

Language Annals 34: 8–17.
Nation, Paul (2001) *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
仁科喜久子（監）鎌田美千子・曹紅荃・歌代崇史・村岡貴子（編）（2012）『日本語学習支援の構築―言語教育・コーパス・システム開発』東京：凡人社.
西尾寅弥（1972）『形容詞の意味・用法の記述的研究』（国立国語研究所報告 44）．東京：秀英出版.
小川典子・李在鎬・横森大輔・土屋智行（2008）「コーパス調査による形容詞の連体形と連用形の頻度」ICJLE.
荻野孝野・小林正博・井佐原均（2003）『日本語動詞の結合価』東京：三省堂.
荻野綱男（編）（2008）『コーパスを利用した国語辞典編集法の研究』文科省科学研究費特定領域研究「日本語コーパス」辞書編集班.
荻野綱男・荻野孝野（2007）「日本語のコロケーション研究の歴史―計量言語学，自然言語処理などを中心に」『日本語学』26(12): 58–70.
小木曽智信・伝康晴（2011）「UniDic2.0：言語資源としての電子化辞書」『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』411–418.
小内一（2010）『てにをは辞典』東京：三省堂.
小野正樹・小林典子・長谷川守寿（2010）『コロケーションで増やす表現』東京：くろしお出版.
小澤俊介・内元清貴・伝康晴（2011）「BCCWJ に基づく中・長単位解析ツール」『特定領域「日本語コーパス」平成 22 年度公開ワークショップ予稿集』331–338.
Palmer, Harold (1938) *A grammar of English words*. London: Longman.
パルデシ，プラシャント・赤瀬川史朗（2012）「BCCWJ を活用した基本動詞ハンドブック作成―コーパスブラウジングシステム NINJAL-LWP の特徴と機能」『特定領域研究「日本語コーパス」 現代日本語書き言葉均衡コーパス完成記念予稿集』205–216.
Pomikálek, Jan and Vít Suchomel (2012) Efficient web crawling for large text corpora. In: Adam Kilgarriff and Serge Sharoff (eds.) *Proceedings of the seventh Web as Corpus Workshop (WAC7)*. Lyon, 2012. (at conference WWW)
斎藤秀三郎（1907）*Saito's class-books of English idiomology*. Tokyo: Kobunsha.
迫田久美子・木下藍子・小西円・李在鎬（2012）「日本語学習者の縦断的会話コーパスの構築と習得研究―3 年間のデータから文法習得の過程を探る―」『日本語教育国際研究大会予稿集（第一分冊）』206.
佐久間尚子・伊集院睦雄・伏見貴夫・辰巳格・田中正之・天野成昭・近藤公久（2005）『NTT データベースシリーズ　日本語の語彙特性　第 3 期（第 8 巻）』書籍 + CD-ROM 版．東京：三省堂.
Sinclair, John (1966) Beginning the studies of lexis. In: C.E. Bazell et al. (eds.), 410–430.
スルダノヴィッチ，イレーナ（2012）「語の共起関係とシラバス―コーパスに準拠した共起表作りの試み―」仁科喜久子（監）鎌田美千子他（編），123–138.
スルダノヴィッチ，イレーナ（2013）「コロケーションとシンタクス―形容詞と名詞のコロケーションを対象に―」『第 4 回コーパス日本語学ワークショップ予稿集』267–274．国立国語研究所.
Srdanović, Irena (2013) Japanese i-adjectives as short and long-word units: Implications to language learning, *PACLING 2013*.
Srdanović, Irena, Tomaž Erjavec and Adam Kilgarriff (2008) A web corpus and word-sketches for Japanese. *Shizen gengo shori (Journal of Natural Language Processing)* 15(2): 137–159.
スルダノヴィッチ, イレーナ・李在鎬（2013）「日本語教育用の形容詞の語彙リストと難易度レベル」『第 3 回コーパス日本語学ワークショップ予稿集』281–290.
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: 59–80.
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（2013）「百億語のコーパスを用いた日本語の語彙・文法情報のプロファイリング」『第 3 回コーパス日本語学ワークショップ予稿集』229–238．国立国語研究所.
Stefanowitsch, Anatol and Stefan Th. Gries (2003) Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2): 209–243.
田野村忠温（2010）「日本語コーパスとコロケーション―辞書記述への応用の可能性―」『言語研究』138: 1–23.
田野村忠温（2012）「日本語のコロケーション」堀正広（編），193–226.
八亀裕美（2008）『日本語形容詞の記述的研究―類型論的視点から』東京：明治書院.
Yamamoto, Tadao (1958) On collocated words in Shakespeare's plays. *Anglica* 3(3): 17–29.
山崎誠（2009）「代表性を有する現代日本語書籍コーパスの構築」『人工知能学会誌』24 (5): 623–631.

Description of Adjective-Noun Collocations Based on Large-Scale Corpora: Towards a Dictionary for Japanese Language Learners

Irena SRDANOVIĆ

University of Ljubljana / Visiting Researcher, NINJAL [–2013.09]

Abstract

Recently, new resources on collocations have emerged, and Japanese language collocation dictionaries have appeared, but there is still a lack of descriptive collocation data based on large-scale contemporary Japanese language corpora. Also, insufficient attention has been given to the systematic treatment of collocations in Japanese language textbooks and teaching materials, which are created mainly by relying on the intuition and experience of language teachers and specialists. Accordingly, the objective of this research is to concentrate on adjective-noun collocations and to describe methods for creation of two resources based on large-scale corpora of contemporary Japanese (BCCWJ and JpTenTen). The first resource is "Adjective-Noun Collocation Data." It uses 500 adjectives as headwords and lists the nouns that combine with each adjective, along with context. A total of 23,247 collocations have already been extracted from the ten-billion-word corpus JpTenTen. The second resource is a "Japanese Language Learner's Dictionary of Adjective-Noun Collocations." It aims at a detailed description of the 25 most frequent basic adjectives, which account for 62% of overall Japanese adjective usage. In this paper, the highly frequent adjective *takai* serves as a model to show how the data from the first resource ("Adjective-Noun Collocation Data") can be used as a basis for the creation of a dictionary for Japanese language learners. The dictionary sorts the modified nouns by difficulty levels and arranges them semantically into lexical maps, putting emphasis on collocations that are difficult for language learners to predict and providing corpus-informed information on register and special usages. Finally, the paper discusses some possible theoretical and practical implications. Once the two resources are complete, they will provide data currently not available for Japanese that can be used for research on lexis and grammar, as well as for the creation of syllabi and language learning materials.

Key words: collocation, corpus, adjectives+nouns, Japanese language education, Japanese linguistics