国立国語研究所学術情報リポジトリ

Construction and Evaluation of "Database of Japanese Compound Verb Examples Based on Web Pages"

メタデータ	言語: jpn
	出版者:
	公開日: 2019-07-25
	キーワード (Ja):
	キーワード (En):
	作成者: 山口, 昌也, YAMAGUCHI, Masaya
	メールアドレス:
	所属:
URL	https://doi.org/10.15084/00002222

国立国語研究所論集 (NINJAL Research Papers) 17: 15–34 (2019)
ISSN: 2186-134X print/2186-1358 online

『Web データに基づく複合動詞用例データベース』の構築と評価

山口昌也

国立国語研究所 研究系 音声言語研究領域

要旨

本論文では，『Web データに基づく複合動詞用例データベース』の構築方法を示し，構築結果を評価する。用例データベースの構築目的は，複合動詞とその構成動詞間の関係を分析することとし，複合動詞の用例，語構成，格解析結果に加え，構成動詞の用例，格解析結果を収録する。本構築手法では，必要な量の用例を確保するために個々の動詞専用の Web コーパスを作成し，Web コーパスの構築量を削減するために漸進的に用例データベースを構築する。この際，用例データベースに登録する複合動詞は，Web から収集可能な量を基準として，半自動的に収集される。本手法により収集した複合動詞は 3371 語，用例数（中央値）は 1173 例であった。この結果は，岩波国語辞典収録語の約 77.2% をカバーする。本手法の評価として，約 2.1 億語の汎用 Web コーパスから収集した用例集合と比較し，幅広い生起確率を持つ複合動詞 1829 語に対して，用例を 1000 例以上収集できることを確認した。また，汎用コーパスから抽出した用例集合の格要素の分布とのコサイン類似度は，複合動詞 0.878，単一動詞 0.919 となった。この結果は，本手法により収集した用例の分布が，汎用コーパスと類似しており，用例収集の偏りが抑制されていることを示唆する *。

キーワード：日本語複合動詞，用例データベース，Web コーパス

1. はじめに

本論文では，『Web データに基づく複合動詞用例データベース』[1]（以後，用例 DB）の構築方法を示すとともに，構築結果を評価する。この用例 DB は，収集した複合動詞のリスト，構成情報とともに，2013 年より Web 上に公開されている。

用例 DB を構築する目的は，日本語の複合動詞と，それを構成する動詞（以後，「構成動詞」と表記）との意味的な関係を分析するための基礎データとすることである。対象とするのは，「探し求める」のような「動詞（連用形）＋動詞」タイプの複合動詞である。

複合動詞と構成動詞との間の意味的・統語的関係については，従来より，言語学，日本語学を中心として，多くの理論的な研究がなされてきた（寺村 1969，長嶋 1976，山本 1984，影山 1993，松本 1998，姫野 1999，由本 2005，石井 2007，陳・松本 2018）。

言語学では，影山（1993）が，語の形成過程の違いにより複合動詞を 2 種類に分類し，それぞれの意味構造，構文構造を分析している。さらに，影山・由本（1997），由本（2005）では，語彙概念構造（Lexical Conceptual Structure，以後，LCS と表記）を用いて，構成動詞と複合動詞の意味構造を記述し，複合動詞と構成動詞との意味的・項構造上の関係を示している。

* 本稿は国立国語研究所 萌芽・発掘型共同研究プロジェクト「文脈情報に基づく複合的言語要素の合成的意味記述に関する研究」（プロジェクトリーダー：山口昌也）の研究成果である。
[1] https://csd.ninjal.ac.jp/comp/

日本語学では，構成動詞に意味的・統語的なカテゴリを与え，複合動詞を類型化した上で，詳細な意味記述を加える研究が主流である。例えば，姫野（1999）では，語彙的複合動詞に関して，構成動詞間の意味的な関係と構成動詞の意味カテゴリに基づいて，複合動詞を類型化している。その上で，複合動詞の格構造，構成動詞と複合動詞の自他の対応関係の他，定性的な意味分析を行っている。

これらの研究では，用例を用いてはいるものの，内省に基づいた分析が主体である。したがって，理論を詳細化したり，自然言語処理に適用したりする場合，大量の用例に基づいた，より客観的な検証が必要になる。特に，複合動詞と構成動詞との格支配の対応関係（山本 1984）や LCS の対応関係（由本 2005）のように意味的な関係を記述する場合，双方の動詞の格パターン，項構造，取りうる格要素といった格関連情報が必要であり，大量の格解析済みの用例が必要不可欠となる。

そこで，本研究では，複合動詞とその構成動詞間の格関連情報の関係分析を目的とした用例データベースを構築する。用例 DB の利用目的を鑑み，すべての用例に格解析結果を付与しておく。用例の収集は，Web から行う。これは，収集できる用例の多様性，量，構築コスト，さらに現在一般に流通している資料であるということを勘案している。このような用例データベースを構築する上で解決すべき課題として，（1）複合動詞のリストの整備，（2）生起確率の低い語の用例収集を挙げる。

まず，（1）の複合動詞のリストについては，国語辞典，形態素解析辞書などを始めとした辞書類，言語研究用の資料（野村・石井 1987，姫野 1999，石井 2007，神崎 2012），日本語教育用の辞書･データベース（Tagashira and Hoff 1986，山下 2007）を利用することが考えられる。しかし，多くの資料が国語辞典の登録語を参考にしており，登録語の選択は辞書の編集方針に依存する [2]。

そのため，網羅性，現在流通している資料への適合性といった点で問題がある。その点，『複合動詞資料集』（野村・石井 1987）は，国語辞典のほか，文学作品や雑誌の語彙調査に基づいて登録語を決定しているが，構築時期が 1980 年代と古い。この他，コーパスから自動的に複合動詞のリストを作成する試み（Breen and Baldwin 2009，加藤・林・伊藤 2011）もなされているが，本研究で対象とする種類の複合動詞を自動的に抽出できていない。以上のことから，現在流通している資料で複合動詞を網羅的に調査し，新たに複合動詞のリストを作成する必要があると考える。

次に，（2）の生起確率の低い語の用例収集についてである。本研究の用例 DB の構築目的は，複合動詞と構成動詞との格関連情報の関係を分析することであり，生起確率の高低に関わらず，複合動詞，構成動詞の用例が一定数以上収集されていることが望まれる。しかし，生起確率の低い語ほど用例収集用のコーパスの規模が大きくなる。しかも，（1）のとおり，複合動詞のリストの作成もあわせて行うため，生起確率は未知であり，構築すべきコーパスの規模を見積もるのが難しい。仮に，単一のコーパスで複合動詞リストの作成，用例 DB の構築を行おうとすると，生起確率の低い複合動詞の探索が十分できなかったり，コーパス構築後に用例不足が明らかになる

[2] 例えば，新語であるために未登録になる場合が考えられる。

といったことも起こりうる。したがって，コーパスの規模の拡大を抑制しつつ，複合動詞の探索と，十分な量の用例収集ができる手法が望まれる。

以上の背景のもと，本研究では，個々の動詞ごとに専用コーパスを構築し，漸進的に用例 DB を構築する手法を用いた。専用コーパスの作成には, Web 検索エンジンを使用した Sharoff（2006）の手法を応用し，収集目標量を設定した用例収集を目指す。また，漸進的な用例 DB の構築方法として，（i）複合動詞の構成動詞になりやすい単一動詞 v_s を種動詞として専用コーパスを構築，（ii）v_s を構成動詞とする複合動詞を探索，（iii）発見した複合動詞の構成動詞を種動詞として（i）～（iii）を繰り返す，といったプロセスを用いる。このプロセスでは，（i）の単一動詞の専用コーパスを用いて，収集目標量を達成できそうな複合動詞を（ii）で探索することにより，コーパス規模の拡大を抑制しつつ，個々の複合動詞に対しても専用コーパスを用いることにより，十分な量の用例の収集を目指す。

本論文の構成は，次のとおりである。まず，2 節で用例 DB の設計を行い，3 節で漸進的な用例 DB 構築の手法を示す。さらに，4 節では，本手法の評価として，（a）収録した複合動詞・単一動詞の量的・語彙的な性質，（b）収集した用例の分布を分析する。最後に 5 節で関連研究に対する本研究の特徴を示したのち，6 節でまとめる。

2. 複合動詞用例データベースの設計

本節では用例 DB の設計として，収録する複合動詞の種類，用例 DB への収録基準，用例の収集目標量を定義したのち，用例 DB の構造と，用例の収集方法について述べる。

2.1 収録する複合動詞の種類

いわゆる「語彙的複合動詞」（影山 1993）を収録の対象とする。影山（1993）では，語の形成過程の違いに基づいて，複合動詞を語彙的複合動詞，統語的複合動詞の 2 種類に分類している。語彙的複合動詞は，語彙部門で形成される。語彙的複合動詞の例として,「使い回す」「出し抜く」「飛び散る」「持ち上げる」が挙げられる。これらは，前項・後項動詞の合成時に，意味的な制限が加わったり，まったく別の意味を持つようになる（例：「出し抜く」）。このように，語彙的複合動詞の辞書は, 構成動詞の辞書から合成的に記述できず, 構成動詞とは別に記述する必要がある。

一方，統語的複合動詞は統語部門で形成される。その前項・後項動詞は,「食べ始める」（＝食べるのを始める）「使い慣れる」（＝使うのに慣れる）といったように補文的な関係を持ち,「意味関係は完全に透明かつ合成的」（影山 1993）であるとされている。このように，統語的複合動詞は，前項動詞と複合動詞との間の構文的関係と意味的な関係が明らかなため，本論文では語彙的複合動詞だけを分析対象とすることにする。なお，以後，特に断りがなければ，語彙的複合動詞は単に「複合動詞」と表記する。

語彙的複合動詞と統語的複合動詞の区別は，代動詞「そうする」によるテスト（影山 1993: 80）を用いる。このテストは「語の中の一部分だけが文中の照応に参加することはできない」という制約を利用する。具体的には，語彙的複合動詞の場合，複合して 1 語となるので，この制約

を満たす。一方，統語的複合動詞の場合，2 語なので，この制約を満たさない。例えば，統語的複合動詞「食べ始める」は「そうし始める」と代動詞で置き換えることができるが，語彙的複合動詞「飛び散る」は「* そうし散る」とすることはできない。

2.2 用例 DB への収録基準と用例の収集目標量

用例 DB へ登録する複合動詞・構成動詞の収録基準は，収集できる用例数によって定める。今回は，複合動詞，単一動詞，それぞれ，50 例以上，100 例以上の用例を収集できる場合，用例 DB へ収録することにした。この量は，複合動詞・構成動詞間の格関連情報の関係分析として，最低限，内省による分析の補助ができるよう設定している。

一方，用例の収集目標量は，共起語ベクトルを用いた語間（本論文の場合は，複合動詞と構成動詞間）の類似性計算（相澤 2008）などの統計的な分析での利用を想定して，設定する。今回は，収集目標量を複合動詞 1000 例，単一動詞 [3] 5000 例とした。

2.3 用例 DB の構造

用例 DB は，収録語リストと，収録語リストに登録されている語の用例集合からなる。図 1 に例を示す。

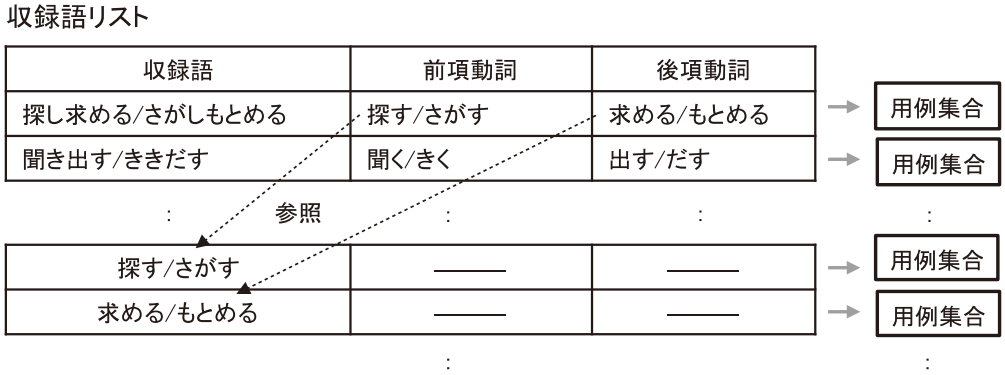

図 1　用例 DB の構造

収録語リストは，用例 DB に収録されている語の情報を格納する。具体的には，語の表記と読み，用例集合へのリンクである。複合動詞の場合，これらに加えて，語構成の情報として，前項動詞，後項動詞へのリンクが含まれる。

それぞれの語は，独立した用例集合を持つ。個々の用例には，格解析結果，出典の URL が付与されているものとする。

[3] 複合動詞が構成動詞となる場合もあるので（例：「引っ張り上げる」の前項動詞「引っ張る」），単一の動詞であることを明記する場合は，「単一動詞」と表記する。

2.4 用例の収集方法

用例収集における基本的な方針は，(i) 収録語ごとに Web データから専用コーパスを構築すること，(ii) Web における生起確率の高低に関わらず，一定量以上の用例を収集することである。この際，言語研究に用いられることを鑑み，収集された用例集合に偏りが生じないよう配慮する。

Web からコーパスを構築する手法自体は，Web を巨大な言語資源として言語研究に利用する流れ（Kilgarriff and Grefenstette 2003）の中で，専門用語を収集することを目的とした特殊コーパスの構築法（Baroni and Bernardini 2004），言語研究用を指向した汎用コーパスの構築法（Baroni et al. 2006）が提案された。構築されたコーパスの規模は数億から 200 億語であり，イギリスの商用のコーパス検索システム Sketch Engine（Kilgarriff et al. 2004）に搭載された。この中には日本語のコーパスとして，約 4 億語の JpWaC（スルダノヴィッチ・仁科 2008），約 103 億語の jpTenTen（スルダノヴィッチら 2013）も含まれる。日本でも今井らが Baroni and Bernardini（2004）を参考に日本語教育用の汎用コーパス「筑波ウェブコーパス」（今井・赤瀬川・プラシャント 2013）を公開している。規模は約 5 億語である。また，浅原ら（2014）は，約 100 億語規模を目指した『国語研日本語ウェブコーパス』を構築している。

Baroni and Bernardini（2004），Baroni et al.（2006）やそれに類する手法の特徴は，目的に応じたキーワードを Web 検索エンジンに与え，取得した URL をもとにして，Web ページを収集することである。Baroni and Bernardini（2004）のように専門用語を収集する場合は，専門用語が含まれるドメインを代表するキーワード集合を用意する。一方，汎用コーパスの場合は，無作為に Web ページが抽出されるよう，ランダムなキーワード集合を用意する。

それに対して，本研究では，指定された語の用例を収集するための専用コーパスを構築する（詳細は，3.2 節を参照のこと）。ただし，収集される用例集合は，汎用コーパスのように無作為に Web ページが抽出され，ひいては，Web 上のテキストの性質が反映されていることが理想である。そのため，Web 検索エンジンに与えるキーワードとして，対象語とランダムなキーワードを組み合わせて利用することにする。

3. 複合動詞用例データベースの構築方法

3.1 概要

構築方法の概要を図 2 に示す。構築のための処理は，大きく分けて，(a) Web コーパス構築，(b) 複合動詞の収集，(c) 用例抽出・格解析，の三つからなる。これらの処理を再帰的に繰り返すことにより，複合動詞を探索的に収集しつつ，複合動詞・単一動詞の用例 DB を漸進的に構築していく。漸進的な構築法を用いたのは，1 節で説明したように，コーパスの規模の拡大を抑制しつつ，複合動詞の探索と，十分な量の用例収集を可能にするためである。この過程を図 2 と対応付けて説明すると，次のようになる。この後の 3.2 節から 3.4 節では，図 2 の (a) (b) (c) の三つの処理を詳しく説明する。

(1) 複合動詞の構成動詞として頻出する単一動詞 v_s を「種」動詞として，v_s の Web コーパス

を構築する。

(2) v_s の Web コーパスから v_s の用例抽出・格解析を行い，用例 DB に追加する。

(3) v_s の Web コーパスから v_s を構成動詞として持つ複合動詞候補リストを機械的に作成する。さらに，複合動詞候補リストから複合動詞 v_c を人手で抽出する。

(4) v_c の Web コーパスを構築し，用例抽出・格解析を行う。用例 DB への収録基準を満たす場合は，用例 DB へ追加する。さらに，v_c の構成動詞が用例 DB に未登録であれば，その構成動詞を種動詞として，再帰的に用例 DB を構築する。

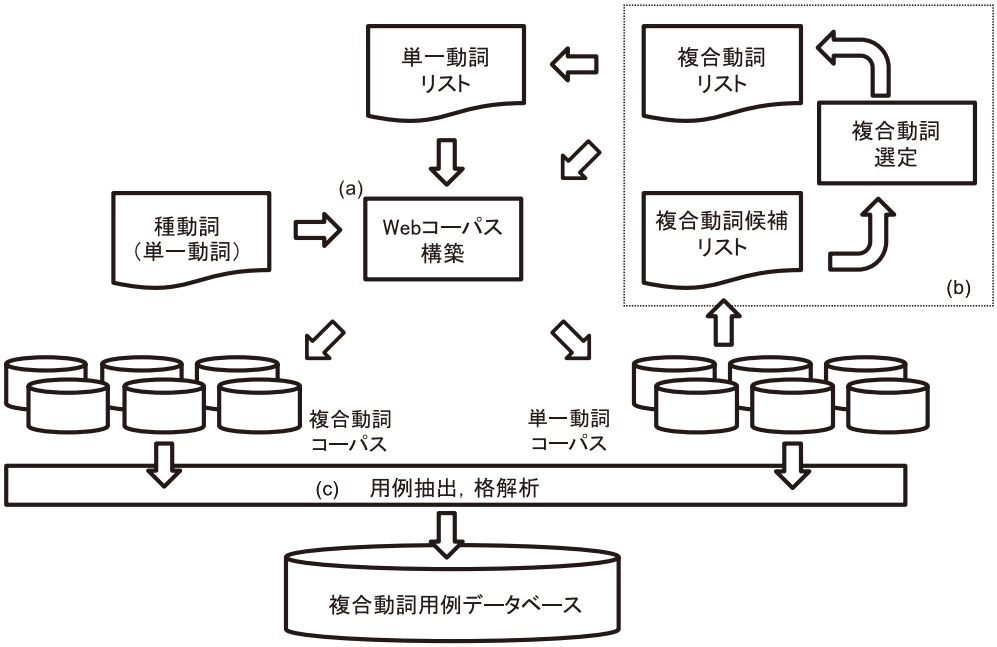

図 2 複合動詞用例データベース構築の流れ

3.2 Web コーパスの構築

ここでは，指定された語 w 専用の Web コーパスを構築するための具体的な方法を示す。この方法は，Sharoff（2006）の手法に基づいている。ただし，2.4 節で述べたとおり，Web 検索エンジンに与えるキーワードとして，ランダムキーワードと w を組み合わせて，利用する。収集する Web ページの量を n_{page} ページ，1 ページあたり 1 例以上の用例が含まれているとすると，理想的には n_{page} 個以上の用例が収集できる。

(1) キーワード集合として，名詞 3000 語を用意した。キーワードの選定基準は，汎用，かつ，多用される名詞とした。これは，広範なサイトから Web ページを収集できるように意図したものである。今回は，「CD- 毎日新聞 2005 データ集」（毎日新聞社 2005）の出現頻度

2000 以上の名詞（ただし，固有名詞，数詞は除く）から，記事ジャンルのエントロピーで上位 3000 語を用いた。なお, 語 w のエントロピー $H(w)$ は次式で定義する。G は記事ジャンルの集合，$p_g(w)$ は語 w が記事ジャンル g に出現する確率である。

$$H(w) = -\sum_{g \in G} p_g(w) \log p_g(w)$$

（2） キーワード集合から, 無作為に n_{key} 個抽出し, Web 検索エンジン[4] で w との AND 検索を行う。なお，n_{key} は単一動詞の場合 2，複合動詞の場合 1 とした。これは，一般的に，複合動詞は単一動詞よりも検索結果数が少なく，条件が多いと，検索結果を得られなくなるためである。

（3） 検索した URL[5] から, 最大 10 個を無作為に抽出する。ただし, ダウンロード済みの URL は, 重複して抽出しない。

（4） 抽出した URL の Web ページをすべてダウンロードする。

（5） n_{page} ページが収集されるまで，（2）～（4）を繰り返す。ただし，この過程で取得可能な Web ページを収集しつくし，永久ループに陥ってしまう場合がある。そのため，検索回数の上限を通常の 2 倍，つまり，$2n_{page}/10$ に制限する[6]。

（6） 収集したページは，各種の正規化[7] を行った上で，形態素解析をしておく。なお，形態素解析システムには，JUMAN Ver.6.0[8] を用いた。

3.3 複合動詞の収集

用例 DB に収録する複合動詞は，一定量以上の用例が収集できる語彙的複合動詞である。一定量以上の用例が収集できるかは，図 2 の（a）Web コーパス構築と（c）用例抽出・格解析を行わないと確定しないため，図 2 全体を包含した形で，複合動詞の収集方法を説明することにする。

（1） 複合動詞の構成動詞として頻出する単一動詞 v_s を「種」動詞として用意する。今回は,『複合動詞資料集』（野村・石井 1987）の調査結果から，複合動詞の構成要素として多用される動詞上位 10 語[9] を種動詞とした。

（2） 3.2 節の方法で v_s に対して, 2 種類のコーパスを構築する。一つは v_s の連用形収集用のコーパスである。このコーパスからは, v_s を前項とする複合動詞候補を収集する。もう一つは, v_s の終止形収集用のコーパスである。このコーパスは，v_s を後項とする複合動詞候補を収集することに加え, v_s の用例収集にも用いる。例えば, 種動詞「取る」の場合,「取る」と「取り」を収集するためのコーパス $C_{取る}$，$C_{取り}$ を作ることになる。収集する Web ページ数は,

[4] Bing Search API（https://datamarket.azure.com/dataset/bing/searchweb）を使用した。
[5] 今回使用した Bing Search API の場合，検索 1 回あたり最大 50 個の URL を取得することができる。
[6] 10 で除算しているのは，1 回の検索あたり，最大 10 ページ取得できるためである。
[7] 文字コードの UTF-8 への統一, HTML タグの除去,（いわゆる）半角文字から全角文字への変換などを行った。
[8] 京都大学黒橋・河原研究室（http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN）
[9]「出す，合う，切る，掛ける，込む，取る，過ぎる，付ける，引く，上げる」の 10 語を用いた。

用例の収集目標量（2.2 節参照）を勘案し，それぞれ 5000 ページとした。

(3) v_s を含む「動詞の連続」を機械的に抽出し，複合動詞候補リストに加える。また，それぞれの複合動詞候補の出現ページ数を計測しておく。例えば，$C_{取る}$ からは「動詞（連用形）＋取る」を抽出し，$C_{取り}$ からは「取り＋動詞」を抽出する。なお，「動詞の連続」とは基本的に「動詞（連用形）＋動詞」だが，形態素解析システムの辞書には，1 形態素として登録されている複合動詞も含まれる。したがって，v_s を先頭，もしくは，末尾に含む動詞も複合動詞候補とする。

(4) 5 ページ以上の Web ページで出現する複合動詞候補の中から，人手で語彙的複合動詞 v_c を選択する。

(5) 選択した複合動詞 v_c 用の Web コーパスを 3.2 節の方法で構築する。Web エンジンへの検索キー（3.2 節の（2））には，v_c の終止形を含める。収集する Web ページ数は，用例の収集目標量（2.2 節参照）を勘案し，2000 ページとする。

(6) 構築された Web コーパスから用例抽出・格解析を行う（3.4 節参照）。

(7) （6）の結果から，一定量以上の用例を収集できた語を用例 DB に登録する（2.2 節で述べたとおり，複合動詞が 50 例以上，単一動詞が 100 例以上の場合）。

(8) 用例 DB に登録した複合動詞の構成動詞が種動詞となっていない場合，その構成動詞を種動詞として再帰的に（2）～（7）を繰り返す。種動詞がすべてなくなった段階で，収録語の選定は終了する。例えば，複合動詞「取り入れる」「選び取る」を登録した場合，「入れる」「選ぶ」が新たに種動詞となる。

3.4 用例抽出・格解析

ここでは，用例の抽出と格解析について述べる。

用例の抽出は，収録対象の動詞 w を含む「文」を単位とする。「文」の認定には，句読点，および，空白，改行を文区切り文字とする，簡易な方法を用いる。抽出された文は，すべて格解析[10]する。格解析には，KNP Ver.3.1[11] を用いた。なお，格解析の際，形態素解析辞書に登録されていない複合動詞は二つの動詞に分割され，格解析結果も動詞ごとに出力される。そのため，両者の結果を統合し，二つの動詞のいずれかに依存する格要素を複合動詞の格要素としている。

このように抽出・格解析された用例は，用例 DB の利用目的に合致するよう，一定の条件を満たしたものだけを集計し，w を用例 DB へ登録することの可否を判定する（3.3 節の手順（7））。条件は，次の四つである。

一つ目の条件は，文長を 4 文字以上 100 文字以下に限定する。これは，文長が短すぎる場合，格関連情報を持たず，用例として不適切となるためである。逆に長すぎると，格解析の処理時間の増大や，商品データなどの単なるデータ列が混入するなどの問題が発生する。二つ目の条件は，

[10] この際，抽出された用例を再度形態素解析しているが，格解析の精度を向上させるために，複数解を出力するようにしている。
[11] 京都大学黒橋・河原研究室（http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP）

w に関して，ヴォイスに関わる受動，使役，可能表現を含まないことである。これは，格要素を比較する際に格の曖昧性を排除するためである。三つ目の条件は，w に関して，一つ以上の格要素を持つということである。四つ目の条件は，同一文の重複登録を許さないことである。

4. 構築結果の評価

4.1 収録語数，用例数

まず，本手法により構築した用例 DB に関して，登録した動詞の語数とその用例数を評価する。

用例 DB に登録した複合動詞，単一動詞の語数，用例数（中央値），出現ページ数（中央値）を表 1 に示す。「出現ページ数」とは，用例が出現した Web ページの異なりである。用例数の目標収集量の点からすると，1000 例以上収集できた複合動詞は 1829 語で，全体の約 54% である。また，5000 例以上収集できた単一動詞は 595 語で全体の約 63%，1000 例以上だと 838 語で全体の約 89% となった。特定のページに用例が集中していないかを確認するために，1 ページあたりの平均用例数（分散）を求めると，複合動詞は 1.3 例 / ページ（0.13），単一動詞は 2.7 例 / ページ（6.8）であった。登録した複合動詞のうち，出現ページ数の上位 20 語を次に示す。

思い込む　言い切る　書き換える　引っ越す　突っ込む　書き込む　買い換える
乗り換える　巻き込む　切り替える　持ち込む　支払う　持ち出す　踏み入れる
詰め込む　割り当てる　見つける　取り付ける　言い出す　知り合う

表 1　用例 DB に登録した複合動詞と単一動詞

	登録語数	用例数（中央値）	出現ページ数（中央値）
複合動詞	3371	1173	946
単一動詞	936	5943	3182.5

次に，収録語の用例数の分布を見るために，複合動詞，構成動詞の用例数のヒストグラムを図 3，4 に示す。横軸は用例数，縦軸は動詞の異なりである。また，用例の出現ページ数のヒストグラムを図 5，6 に示す。横軸は出現ページ数，縦軸は動詞の異なりである。

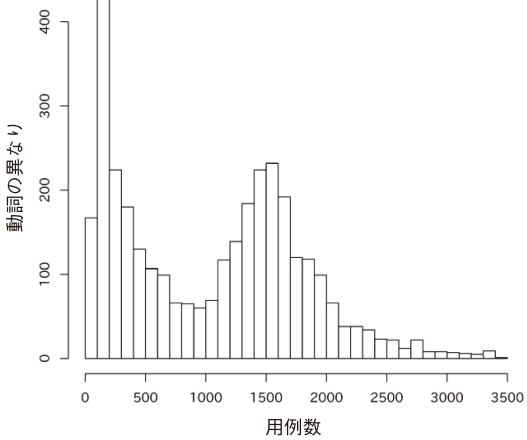

図 3　複合動詞用例数

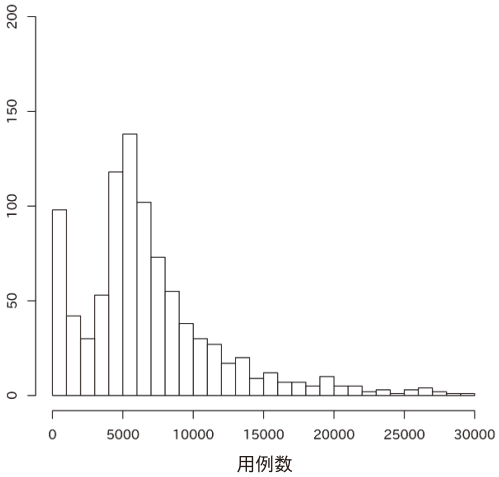

図 4　単一動詞用例数

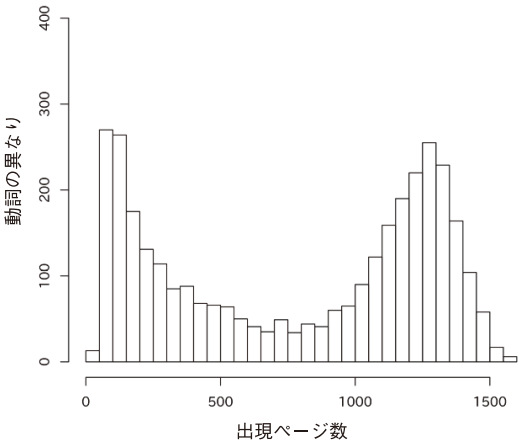

図 5　複合動詞出現ページ数

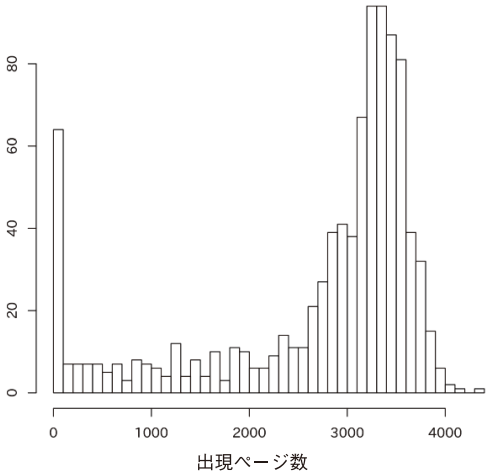

図 6　単一動詞出現ページ数

これらの分布の特徴的な点は，ピークが二つあることである。複合動詞の場合，図 5 で顕著なように，高頻度区間のピーク（1250 ページ以上 1300 ページ未満），低頻度区間のピーク（50 ページ以上 100 ページ未満）の 2 箇所にピークが存在する。高頻度の複合動詞群は，広範囲の Web ページで用いられていることから，一般的な使用が定着している複合動詞群と考えられる。低頻度の複合動詞群は，限られた Web ページでしか利用されていない。実際，Web コーパス構築の段階で Web ページの取得自体ができておらず，0 ページ以上 250 ページ未満の複合動詞の Web コーパスの収集ページ数は，平均 312.9 ページであった。したがって，「新造」の複合動詞（石井 2007）や専門用語などの可能性が考えられる。以上は量的な分析だが，質的な分析は，4.2 節で国語辞典との比較により行う。

一方, 単一動詞（図 6）でも, 高頻度区間（3200 ページ以上 3400 ページ未満）と低頻度区間（0 ページ以上 100 ページ未満）にピークがある。ただし，低頻度区間の単一動詞群（64 語）には，形態素解析上の問題で用例数 0 となった 61 語も含まれる。したがって，実質的には，単一動詞の分布は単一のピークを持つ分布である。なお，低頻度区間の単一動詞群は，古語など，現代では単独での使用が少ない語である。例として,「古す」（ふるす),「躙る」（にじる),「伸べる」（のべる),「交う」（かう）を挙げる。

4.2 既存辞書との比較

本手法により収集された複合動詞，されなかった複合動詞の質的な評価をするために，既存の辞書に登録されている語義的複合動詞と比較する。比較対象は, 岩波国語辞典第五版（岩波書店・三菱電機 2010）（以後, 岩波）と JUMAN Ver.6.0 の辞書（以後, JUMAN）とした。比較の観点は，次の 2 点である。

観点 1： 既存辞書の登録語は語彙的に重要な語であると仮定し，それが収集された用例数にどのように反映されているか調べる

観点 2： 既存辞書の登録語であるにも関わらず，収集できなかった複合動詞を調べ，その原因を探る

基本的なデータとして，用例 DB との比較結果を表 2 に示す [12]。収集した複合動詞は岩波の収録語の約 77.2% をカバーし，規模は岩波の収録語の約 2.1 倍である。また，JUMAN の収録語の約 97.1% をカバーし，規模は JUMAN の収録語の約 3.1 倍である。

表 2　既存の辞書との比較結果（語数と例）

	用例 DB のみ	比較対象のみ	両方
岩波	2113 書き込む，見開く 植え替える，騙し取る	371 盛り潰す，蹴倒す ひっ捕らえる	1258
JUMAN	2310 乗り換える，飛び込む 取り上げる，煮込む	31 ずりさがる，打ち拉ぐ 済し崩す，聞き古す	1055

まず,「観点 1」として，用例 DB と比較対象の辞書の両方に登録されている語について，用例数の分布を見てみる。図 7 中の × は，岩波と用例 DB の両方に収録されている語（以後,「岩波重複語」）の用例数の分布を図 5 に重ねてプロットしたものである。また，図 8 には，図 7 中の × をヒストグラムで表示し，そこに JUMAN と用例 DB の両方に収録されている語の用例数の分布を * でプロットしている。図 8 を見ると，用例 DB と重複する語については，岩波と

[12]「語」の単位は, 資料ごとに異なるが, 表中の「用例 DB のみ」「両方」欄では用例 DB の基準,「比較対象のみ」欄では比較対象の辞書の基準で計測している。具体的には，岩波では 1 見出しを 1 語とし，JUMAN では同一の代表表記を持つ場合は同一の語として計測した。

JUMAN では大きな差がない。また，用例 DB に登録されていない語は岩波のほうが約 12 倍以上多い（表 2）。そこで，以後の考察では，岩波の結果について詳しく見ていくことにする。

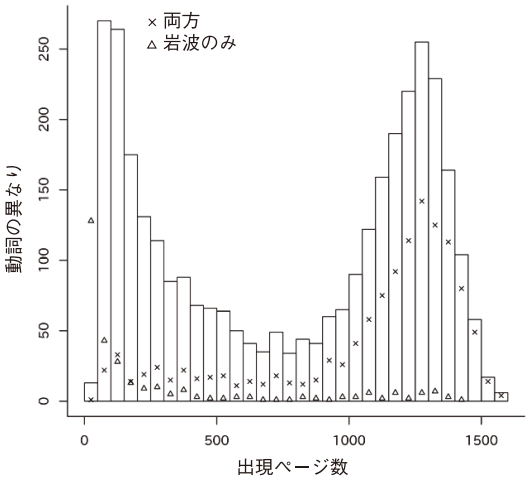

図 7　構築結果と岩波との比較

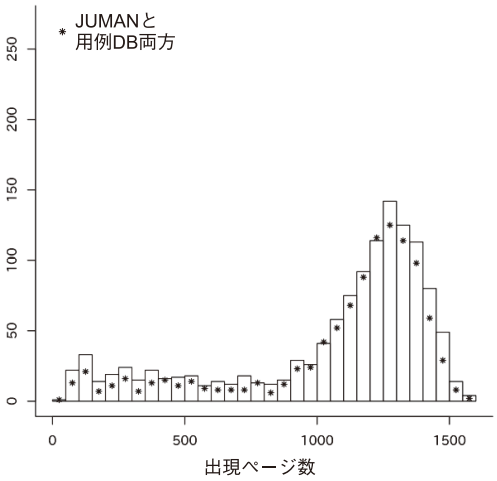

図 8　岩波と JUMAN との比較

岩波重複語の用例数の分布（図 7 の×）は，用例 DB の複合動詞の用例数の分布のうち，高頻度の複合動詞群と重なる部分が多い。実際に，岩波重複語の用例数は中央値 1209 ページで，その約 72.0% が用例 DB において出現ページ数 1000 ページ以上の動詞である。この結果は，用例 DB の高頻度の複合動詞群が使用の定着した語を多く含んでいることの左証となる。

次に，「観点 2」として，岩波収録語のうち，用例 DB に登録されなかった複合動詞について見てみる。該当する複合動詞は 371 語あった（表 2）。これらの複合動詞の用例を本手法で収集したところ，371 語のうち，183 語が用例 DB への追加の基準を満たした。これらの語の出現ページ数の分布は，図 7（△の点）のようになった。出現ページ数は平均 434.1 ページ，中央値は 206 ページである。

用例 DB への追加の基準を満たした 183 語の複合動詞が，用例 DB 構築時になぜ登録されなかったかを調査する。ここでは，Web コーパス構築，複合動詞の収集の過程を調査した。その結果，101 例で，複合動詞候補の出現頻度がコーパス構築基準（3.3 節（4））に達していなかった。また，64 例が複合動詞選択時（3.3 節（4））の人手判断ミスであった。

さらに，用例 DB への追加の基準を満たさなかった 188 語について，その原因を見てみると，64.9% が登録規定数（3.3 節（7））の用例を収集できていなかったことによるものだった。また，32.0% が形態素解析の失敗に起因する用例抽出の失敗だった。具体的には，古語を始めとした未知語，および，未登録の音便形の存在により，形態素解析に失敗した。以下に例を示す。本手法では，このような構成動詞を持つ複合動詞は，形態素解析用の辞書に 1 語として登録されていない限り，複合動詞候補（図 2）となりえない。

収集量不足の例：盛り潰す，振り放つ，差し付ける，推し移る，焦れ込む，尋め行く
未知語の例：蹴倒す（「蹴る」の下一段活用），出立つ（「出づ」），凍てつく（「凍てる」）
音便形の例：引っ捕らえる，突っ伏す，ぶっぱなす

4.3 無作為抽出文コーパスとの比較

本節では，用例 DB と「無作為抽出文コーパス」から収集した用例集合とを比較する。ここで言う「無作為抽出文コーパス」（以後，「RS コーパス」と表記）とは，複数の Web ページから無作為に抽出した文から構成されるコーパスである。比較の目的は，用例が偏りなく収集できているか検証することである。比較の観点は，用例数と用例の分布とする。

まず，RS コーパスの構築方法について説明する。RS コーパスの原資料は，用例 DB を構築する過程で作成した単一動詞の Web コーパスから選択することにした[13]。選択の条件は，特定ジャンルへの偏りを防ぐため，収集対象動詞の用例を 1000 例以上収集できた動詞の Web コーパスとした。この条件に適合するすべての Web コーパスから重複ページを除去した Web ページ（約 449 万ページ。重複する文を削除すると，約 16 億 7243 万文）が原資料である。RS コーパスは，この原資料からページあたり最大 3 文をランダムに抽出し，マージしたものである。ただし，抽出した文と同一の文がすでに抽出されていた場合は，抽出をやり直した。なお，ここでの「文」の定義は 3.4 節とほぼ同様である。用例 DB 構築時と異なる点は，文長の下限を 4 文字から 10 文字としたことである。制限を厳しくしたのは，RS コーパスでは抽出対象の動詞の制約がなく，格要素を含まない文字列が抽出されやすくなるためである。以上の処理の結果，RS コーパスは，約 1330 万文，約 2.1 億形態素を収録するコーパスとなった。

一つ目の比較の観点として，用例数について見てみよう。ここでは，用例 DB で用例数が 1000 例以上の複合動詞 1829 語を対象として，3 節の手法で RS コーパスから用例を抽出する。

図 9 は，抽出した用例数のヒストグラムである。用例数の中央値・平均値はそれぞれ 102 例・315.4 例，用例数の範囲は 0 〜 8492 例であった。用例数の上位，下位 30 語を次に示す（括弧内の数字は用例数）。なお，単一動詞に対しても，同様の条件で用例を抽出した結果，用例数の中央値・平均値はそれぞれ 1837 例・2976.8 例となった。用例数の範囲は，53 〜 9747 例である。

上位 30 語：見つける (8492)，見つかる (8311)，思い出す (8146)，繰り返す (6608)，落ち着く (6177)，取り組む (6176)，出会う (6019)，見かける (5859)，受け入れる (5441)，出かける (4828)，受

[13] 本来，RS コーパスの原資料は用例 DB の原資料とは別に作成すべきだが，その影響は軽微であるため，今回の方法を用いた。具体的な影響の度合いとしては，動詞 v 用に構築した Web コーパスを RS コーパスの原資料に含めると，新規に構築した原資料の場合と比較して，RS コーパス中の v の用例数の増加量は，高々約 0.14 例である。内訳は次のとおりである。まず，収集した動詞のうち，最も用例数が多いのは，単一動詞の「思う」で 171358 例であった。したがって，「思う」の用例を RS コーパスから取得する場合に，増加量が最大となる。このときの増加量は，$171358 \times 2 \times a/b$ (≈ 0.14) となる。a は RS コーパスに含まれる文数（約 1330 万文），b は原資料に含まれる文数（約 16 億 7243 万文）である。2 を乗じているのは，3.3 節（2）で示したとおり，単一動詞は終止形，連用形で 2 種類の Web コーパスを作るためである。ただし，連用形の Web コーパスは複合動詞探索用にしか用いず，そこに含まれる用例数は計算していないので，終止形の Web コーパスの用例数で代用した。

け取る (4743)，申し上げる (4651)，振り返る (4600)，取り入れる (4544)，生み出す (4425)，取り付ける (4403)，盛り上がる (4270)，取り上げる (4261)，付き合う (4179)，支払う (4115)，書き込む (3769)，見直す (3653)，組み合わせる (3508)，取り戻す (3431)，見つめる (3345)，似合う (3109)，引き出す (3092)，取り出す (3057)，引き起こす (2866)，乗り越える (2770)

下位 30 語：かいくぐる (0)，飲み交わす (0)，吹き降ろす (4)，守り勝つ (5)，混ぜ合わす (6)，消し込む (6)，弾き語る (6)，弾き込む (6)，這い寄る (6)，なめ回す (7)，押し負ける (7)，撃ち負ける (7)，晒し上げる (7)，死に逝く (7)，叩き斬る (7)，鋳込む (7)，すり混ぜる (8)，掛け替える (8)，差し障る (8)，作り分ける (8)，産み分ける (8)，使い潰す (8)，思い合う (8)，勝ち切る (8)，振り遅れる (8)，叩き直す (8)，張り倒す (8)，投げ下ろす (8)，舞い飛ぶ (8)，こね回す (9)

以上の結果から，幅広い生起確率の複合動詞の用例を，収集目標量の 1000 例以上収集できていることがわかる。さらに強調したい点は，生起確率が低い複合動詞の中にも，語彙的に重要な語が含まれていることである。これは, 岩波国語辞典に収録されている複合動詞（図 9 の○の点）が，用例数 50 例未満，50 以上 100 未満の区間でも，それぞれ 34.2%，47.9% 含まれていることからわかる。

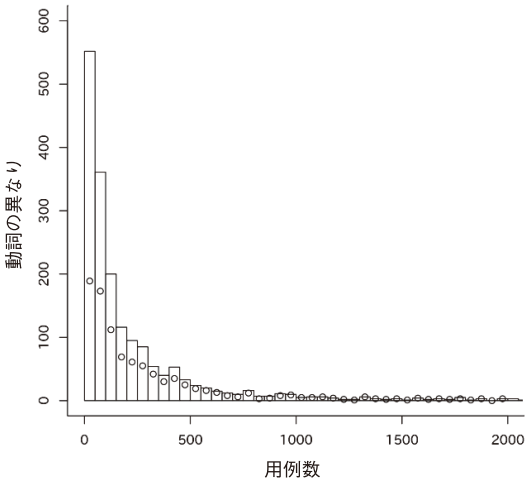

図 9　RS コーパスにおける複合動詞の用例数

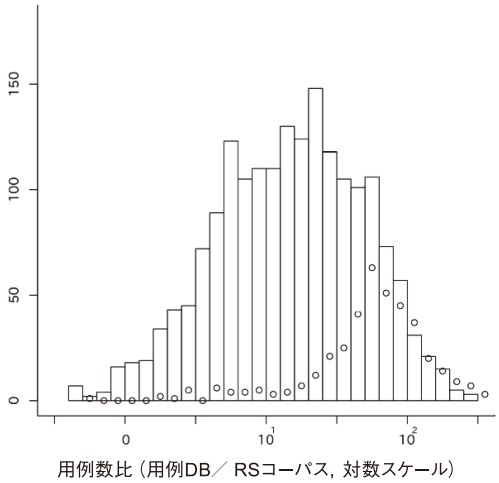

図 10　用例数比（用例 DB/RS コーパス）

次に，単一のコーパスにより用例 DB を構築した場合，どれくらいの規模のコーパスが必要かを把握するために，用例 DB と RS コーパスの用例数比を求めた。用例数比のヒストグラムを図 10 に示す。対象とした動詞は，用例 DB で用例数 1000 例以上の複合動詞である。横軸は，用例数比（用例 DB の用例数 /RS コーパスの用例数，対数スケール）である。用例数比の中央値・平均値は，それぞれ 15.5・27.6 であった [14]。この結果から，用例数比が中央値以下の複合動詞の用

[14] 用例数が 0 だった 2 動詞は除いて計測した。

例 DB を構築するには，生起確率が RS コーパスと同一だと仮定した場合，RS コーパスの規模の 15.5 倍（≈ 32.6 億語）のコーパスを構築する必要があると考えられる。用例 DB での用例数が 500 以上 1000 例未満の複合動詞になると，用例数比の中央値・平均値はそれぞれ 61.5・79.8 となり，用例数比がさらに増大する。

最後に，二つ目の観点である，格要素の分布の類似性に関して，用例 DB と RS コーパスとを比較する。ここでは，比較の尺度として，コサイン類似度 $\cos(\mathbf{w_1},\mathbf{w_2})$ を用いた。

$$\cos(\mathbf{w_1},\ \mathbf{w_2}) = \frac{\mathbf{w_1} \bullet \mathbf{w_2}}{|\mathbf{w_1}||\mathbf{w_2}|}$$

$\mathbf{w_1}, \mathbf{w_2}$ は，それぞれ動詞 w_1, w_2 に依存する格要素[15] の頻度ベクトルである。頻度ベクトルの要素は，格マーカと格要素のペアの出現ページ数とした。頻度ベクトルの計算には，3.4 節の格解析によって得られた結果を用いた。なお，解析誤りなどのノイズを避けるため，頻度 3 未満のベクトル要素は 0 としている。また，品詞の情報は捨象した。

今回は，用例 DB，RS コーパスの両方で，用例数が 1000 例以上の複合動詞，単一動詞に対して，コサイン類似度を計算した。コサイン類似度の平均，分散，対象動詞の数を図 11，12 に示す。

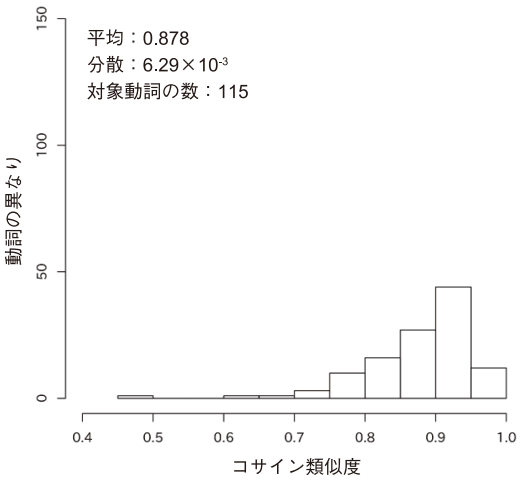

図 11　複合動詞のコサイン類似度

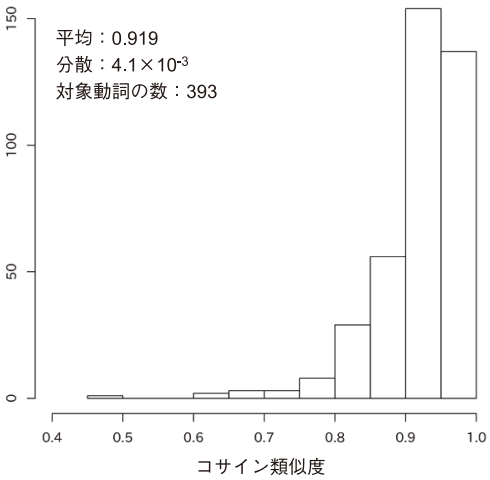

図 12　単一動詞のコサイン類似度

この結果のとおり，用例 DB と RS コーパスの用例集合間のコサイン類似度は，複合動詞 0.878，単一動詞 0.919 となった。これは，本手法により収集した用例集合が RS コーパスに類似しており，用例収集に関しての偏りが抑制されていることを示唆するものである。本実験では，用例 DB・RS コーパス双方の用例数が 1000 例以上の動詞のみでしか検証できていないが，Web ページ取得時（3.2 節の（2））に，十分な量の Web ページを偏りなく収集できれば，本実験と同等の類似度を持つ用例集合を構築できると考えられる。

[15] 格マーカが格助詞の格要素のみを対象とした。

ただし，図 11，12 には，コサイン類似度が低い動詞も存在する。コサイン類似度が 0.7 以下の動詞の格要素を分析したところ，Web ページ取得時に特定の活用形（本研究の場合は，基本形）を検索キーとすることによる，格要素の分布の変化が観察された。最も顕著なのが，単一動詞「交える」である。この動詞のコサイン類似度は 0.416 と，図 12 中の単一動詞で最も低い。「交える」のヲ格要素について見てみると，RS コーパスでは，連用形で使用されることの多い「（事例を）交え，～する」といった語義の格要素が上位 10 語中 9 語を占める。それに対して，用例 DB では基本形を検索キーとするため，「（剣などの武器を）交える」といった別の語義の格要素が上位 10 語中 8 語に変化する。同様の傾向は，他の語にも見受けられる。以下に，用例 DB と RS コーパスで生起確率の変動が大きい格要素の例を示す。なお，括弧内の数字は，当該格における生起確率の比である。

用例 DB で生起確率大：「（大は）小を兼ねる」(3.0)「（～で）用が足りる」(17.8)
RS コーパスで生起確率大：「生産が追いつかない」(7.5)「～の合間を縫って」(2.3)

図 11，12 を見る限り，上記の問題の影響は限定的だが，すべての語の用例集合に影響を与えている。また，用例 DB の構築目的である，複合動詞と構成動詞との比較においても，特定の活用形でしか出現しない用法の用例数が減少する可能性がある。この問題の解決は，今後の課題である。

5. 関連研究

本節では，(1）特定目的に特化した用例データベースの構築と，(2）汎用のコーパスを用例データベースとして利用するという観点から関連研究を概観し，本研究の特徴を明確にする。

5.1 特定目的の用例データベースの構築

ここでは，特定の目的に特化した用例データベースの構築に関する研究を，（a）既知の語集合に対して用例を収集する研究，(b）収録対象語を発見しつつ用例収集を行う研究に分けて概観し，本研究と比較する。

まず，前者の研究としては，multiword expression に関連した研究がある。土屋ら（2006）は日本語複合辞用例データベース，Hashimoto and Kawahara（2009）は日本語イディオムの用例データベース，Cook et al.（2008）は動詞と名詞から構成される英語イディオムの用例データベースを構築している。用例収集のための原資料は，それぞれ，毎日新聞（9 年分），Web コーパス（Kawahara and Kurohashi 2006），BNC[16] である。これらの研究は，解釈の曖昧性（字義どおりに解釈，慣用句的に解釈）を解消するための基礎資料収集を目的としており，慣用的か否かの情報がアノテーションされている。複合動詞に特化したデータベースとしては，『複合動詞レキシコン』（神崎 2013）がある。このデータベースは，言語研究，日本語教育での利用を想定しており，

[16] http://www.natcorp.ox.ac.uk/

既存の辞書などから収集した語彙的複合動詞 2735 語に対して，語構成，格パターン，自他・主語の意思性の有無，語義などの辞書的な情報とともに，数例ずつ用例が付与されている。動詞一般では，河原・黒橋（2006）が Web コーパスから格フレーム辞書を自動構築し，格フレームごとに用例がまとめられている。また，動詞の項構造については，LCS による記述の研究（竹内ら 2006，平野ら 2006）が，項構造辞書に対して項構造タグ付き用例の自動付与を試みている。用例収集のための原資料は毎日新聞（13 年分）で，規模は 4 動詞に対して，8385 例である。

収録語を発見しつつ用例収集を行う研究としては，オノマトペの辞書構築に関する研究がある。奥村ら（2003）は，言語学的な研究に基づく音韻パターンからオノマトペ候補を生成し，Web 検索数と Web ページからの用例収集結果から収録語を決定している。また，古武・佐藤（2010）では，収録語自体は既存の辞書などから取得したオノマトペであるが，後接する付属語の規則に基づいて，用例別にサブエントリを生成している。サブエントリごとに行われる用例収集のための原資料は「現代日本語書き言葉均衡コーパス」[17] である。

以上の研究に対して，本研究は（b）のタイプに分類され，特に，奥村ら（2003）に最も近い。奥村ら（2003）に対して，本研究の特徴的な点は，原資料（本研究の場合は Web）の性質を，収集された用例集合の性質に反映させつつ，一定量以上の用例が収集できるようにしている点である。この特徴は Web テキストを対象とした言語研究や自然言語処理において有用であると考える。

5.2 汎用コーパスの利用

2.4 節で述べたとおり，Web コーパスを始めとする，大規模な汎用コーパスが存在する。したがって，複合動詞の認定の問題を別にすれば，既存のコーパスを複合動詞の用例データベースとして，利用することができる。コーパスを利用する環境も整備されてきており，レキシカルプロファイリング型の検索ツール，例えば，前述の Sketch Engine や，「現代日本語書き言葉均衡コーパス」・「筑波ウェブコーパス」で活用されている汎用検索ツール NINJAL-LWP（パルデシ・赤瀬川 2012），『梵天』（浅原ら 2018）を用いれば，用例や格関連情報を容易に抽出することができる。

そこで，汎用コーパスから抽出される用例集合と用例 DB の用例集合とを比較したところ，次のようになった。

- 用例 DB では，4.1 節で示したように，目標収集量に応じた量の用例集合を得ることができる。一方，汎用コーパスから得られる用例集合の量は，汎用コーパスにおける生起確率に依存するため，生起確率が低い語ほど大規模なコーパスを構築する必要がある（4.3 節参照）。
- 用例 DB では，Web 全体のテキストの性質を考慮しつつ，既存の用例集合へ用例を追加することが可能である。汎用コーパスでも，Sharoff（2006）の手法のように，無作為に

[17] https://www.ninjal.ac.jp/corpus_center/bccwj/

Web ページを収集していく手法であれば，用例の追加収集が可能である。ただし，収集したい語だけでなく，コーパス全体の拡張が必要となる。

- 用例 DB では, 個々の語の用例集合は互いに独立している。このことの利点は, 巨大なコーパスを作らずに，語単位の調査・分析ができることである。本論文と同一の収集目標量であれば，1 語あたり，2000Web ページ収集すればよい。これは，特定の語や，語群を限定した言語研究[18] にとって有用な特徴である。一方, 欠点は原資料全体における生起確率が不明なことである。この問題の解決には，4.3 節で構築したような，小規模の汎用コーパスを構築するか，既存のコーパスを利用する必要がある。

6. おわりに

本論文では, 複合動詞, 構成動詞の用例データベースを構築し, 構築結果を評価した。用例データベースに格納される情報は，複合動詞の用例，語構成，格解析結果に加え，複合動詞を構成する動詞の用例，格解析結果である。本手法では，個々の語専用の Web コーパスを構築し，漸進的に用例データベースを構築する。この際，用例データベースに登録する複合動詞は，Web から収集可能な用例の量を基準として，半自動的に収集される。

本手法を用いて用例データベースを構築した結果，複合動詞，単一動詞の登録語数はそれぞれ 3371 語，936 語，用例数（中央値）はそれぞれ 1173，5943 例であった。

構築した用例データベースを評価するため，岩波国語辞典の収録語との比較，および，約 2.1 億語の RS コーパスから収集された用例集合との比較を行い，次の結果を得た。

- 用例 DB に登録された複合動詞は，岩波の収録語の 77.2% をカバーし，規模は約 2.1 倍である。また，岩波重複語の用例数の分布は，用例 DB 中の高頻度の複合動詞群と重なる部分が多い。
- 用例 DB で用例数 1000 例以上の複合動詞を対象に，RS コーパスから用例収集を実施したところ，用例数 0 ～ 8492 例（中央値 102 例）の用例集合を得た。これにより，本手法は幅広い生起確率の動詞に対して，1000 例以上の用例を収集できることがわかった。
- 用例 DB，RS コーパスで用例数 1000 例以上の動詞を対象に，両者の用例集合の格要素の分布を比較したところ, コサイン類似度で複合動詞 0.878, 単一動詞 0.919 となった。これは，本手法により収集した用例の分布が RS コーパスに類似しており，用例収集に関しての偏りが抑制されていることを示唆する。

今後の課題としては，Web コーパス構築時に発生する，活用形による用例の偏りを補正することと，複合動詞の収集における人手判断ミスを防ぐことである。

[18] 例えば，松田（2001）では，後項動詞が「こむ」の複合動詞を研究対象としている。

参照文献

相澤彰子（2008）「大規模テキストコーパスを用いた語の類似度計算に関する研究」『情報処理学会論文誌』49(3): 1426–1436.

浅原正幸・今田水穂・保田祥・小西光・前川喜久雄（2014）「Web を母集団とした超大規模コーパスの開発 収集と組織化」『国立国語研究所論集』7: 1–26.

浅原正幸・河原一哉・大場寧子・前川喜久雄（2018）「『国語研日本語ウェブコーパス』とその検索系『梵天』」『情報処理学会論文誌』59(2): 299–306.

Baroni, Marco and Silvia Bernardini (2004) BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi and Eros Zanchetta (2006) The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3): 209–226.

Breen, James and Timothy Baldwin (2009) Corpus-based extraction of Japanese compound verbs. *Proceedings of the 2009 Australasian Language Technology Workshop (ALTW 2009)*: 35–43.

陳奕廷・松本曜（2018）『日本語語彙的複合動詞の意味と体系―コンストラクション形態論とフレーム意味論』東京：ひつじ書房.

Cook, Paul, Afsaneh Fazly and Suzanne Stevenson (2008) The VNCTokens dataset. *Proceedings of the LREC Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008)*: 19–22.

古武泰樹・佐藤理史（2010）「用例に基づくオノマトペ用法辞書の編纂」『言語処理学会第 16 回年次大会論文集』994–997.

Hashimoto, Chikara and Daisuke Kawahara (2009) Compilation of an idiom example database for supervised idiom identification. *Language Resources and Evaluation* 43(4): 355–384.

姫野昌子（1999）『複合動詞の構造と意味用法』東京：ひつじ書房.

平野徹・飯田龍・藤田篤・乾健太郎・松本裕治（2006）「動詞項構造辞書への大規模用例付与」『自然言語処理』13(3): 113–132.

今井新悟・赤瀬川史朗・プラシャント パルデシ（2013）「筑波ウェブコーパス検索ツール NLT の開発」『第 3 回コーパス日本語学ワークショップ予稿集』199–206.

石井正彦（2007）『現代日本語の複合語形成論』東京：ひつじ書房.

岩波書店・三菱電機（2010）『岩波国語辞典第五版タグ付きコーパス 2004』言語資源協会.

影山太郎（1993）『文法と語形成』東京：ひつじ書房.

影山太郎・由本陽子（1997）『語形成と概念構造』東京：研究社.

加藤恒昭・林良彦・伊藤たかね（2011）「語釈文を用いた複合動詞の特徴分類」『言語処理学会第 17 回年次大会予稿集』568–571.

Kawahara, Daisuke and Sadao Kurohashi (2006) Case frame compilation from the web using high-performance computing. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

河原大輔・黒橋禎夫（2006）「格フレーム辞書の漸次的自動構築」『自然言語処理』12(2): 109–131.

神崎享子（2012）「複合動詞データベース構築のための付与情報」『国立国語研究所論集』3: 1–18.

神崎享子（2013）「『複合動詞レキシコン』ver.1―形態的・統語的・意味的情報付与―」『言語処理学会第 19 回年次大会予稿集』761–764.

Kilgarriff, Adam and Gregory Grefenstette (2003) Introduction to the special issue on the web as corpus. *Computational Linguistics* 29(3): 333–347.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz and David Tugwell (2004) The Sketch Engine. *Proceedings of EURALEX*.

毎日新聞社（2005）「CD- 毎日新聞 2005 データ集」日外アソシエーツ.

松田文子（2001）「コア図式を用いた複合動詞後項『～こむ』の認知意味論的説明」『日本語教育』111: 16–25.

松本曜（1998）「日本語の語彙的複合動詞における動詞の組み合わせ」『言語研究』114: 37–84.

長嶋善郎（1976）「複合動詞の構造」『日本語講座 4 日本語の語彙と表現』東京：大修館書店.

野村雅昭・石井正彦（1987）『複合動詞資料集』（科学研究費補助金 特定研究（1）「言語データの収集と処理の研究」報告書）.

奥村敦史・斎藤豪・奥村学（2003）「Web 上のテキストコーパスを利用したオノマトペ概念辞書の自動構築」『情報処理学会研究報告．自然言語処理研究会報告』2002-NL-154: 63–70.

パルデシ プラシャント・赤瀬川史朗（2012）「レキシカルプロファイリング手法を用いた BCCWJ 検索ツール NINJAL-LWP とその研究事例」『日本言語学会第 144 回予稿集』364–369.

Sharoff, Serge (2006) Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics* 11(4): 435–462.
スルダノヴィッチ エリャヴェッツ 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 エリャヴェッツ 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.
Tagashira, Yoshiko and Jean Hoff (1986) *Handbook of Japanese compound verbs*. Tokyo: The Hokuseido Press.
竹内孔一・乾健太郎・藤田篤（2006）「語彙概念構造に基づく日本語動詞の統語・意味特性の記述」『レキシコンフォーラム』2: 85–120. 東京：ひつじ書房.
寺村秀夫（1969）「活用語尾・助動詞・補助動詞とアスペクト（その一）」『日本語・日本文化』1: 32–48. 大阪外国語大学.
土屋雅稔・宇津呂武仁・松吉俊・佐藤理史・中川聖一（2006）「日本語複合辞用例データベースの作成と分析」『情報処理学会論文誌』47(6): 1728–1741.
由本陽子（2005）『複合動詞・派生動詞の意味と統語』東京：ひつじ書房.
山本清隆（1984）「複合動詞の格支配」『都大論究』21: 32–49.
山下喜代（2007）『日本語教育のための合成語のデータベース構築とその分析』（科学研究費補助金基盤研究（C）研究成果報告書）.

Construction and Evaluation of "Database of Japanese Compound Verb Examples Based on Web Pages"

YAMAGUCHI Masaya

Spoken Language Division, Research Department, NINJAL

Abstract

This paper presents a method of constructing a database of Japanese compound verb examples, and evaluates the database. The objective of constructing this database is to analyze the relationship between Japanese compound and component verbs. Whether to include a compound verb in the database is determined semi-automatically by the number of examples that can be extracted from these Web corpora. The actual database that resulted from this method consists of 3371 compound verbs (median number of examples per verb = 1173). It covers 77.2% of the relevant entry words in the Iwanami Japanese language dictionary. A comparison with a general-purpose Web corpus shows that this method enabled to collect more than 1000 examples for 1829 compound verbs with a wide range of probability of occurrence. The average cosine similarity between the distributions of case-marked elements in the database examples and in those extracted from the Web corpus is 0.878 for compound verbs. Therefore, this result suggests that the bias of examples is controlled.

Key words: Japanese compound verb, example database, web corpus