Web データに基づく複合動詞データベースの構築

山口昌也 (国立国語研究所言語資源研究系) †

Constructing a Japanese Compound Verb Database Based on Web Pages

Masaya YAMAGUCHI (Dept. Corpus Studies, NINJAL)

1 はじめに

本稿では，日本語の複合動詞データベースの構築について述べる。対象とする複合動詞は,「切り倒す」「持ち上げる」といった,「動詞（連用形）＋動詞」タイプの複合動詞である。

本データベースを構築する目的は，複合動詞とそれを構成する動詞（以後，構成動詞）との関係を，大量の用例に基づいて分析することである。本研究で特にターゲットするのは，複合動詞と構成動詞の格要素の対応関係である。

例えば，複合動詞「切り倒す」「打ち破る」のヲ格について，構成動詞との対応関係を見てみる。まず,（E1）「切り倒す」の「木を」の場合は,「切る」（E1a）「倒す」（E1b）ともに適格である。一方,（E2）「打ち破る」の「記録を」の場合は,「打つ」（E2a）が不適格である。複合動詞と構成動詞の関係を調べるためには，以上のようなテストを多数行うことにより,「切り倒す」のヲ格は両方の構成動詞のヲ格と対応関係があり,「打ち破る」のヲ格は「破る」のヲ格とだけ関係があると判断する。

E1 太郎が木を切り倒す　　**E2** 太郎が記録を打ち破る
E1a 太郎が木を切る　　**E2a** * 太郎が記録を打つ
E1b 太郎が木を倒す　　**E2b** 太郎が記録を破る

このような格要素の対応関係分析は，複合動詞と構成動詞間の格支配関係分析（山本 1984 など）や LCS による意味構造の記述（影山 1993，由本 2005 など）といった，より複雑な分析の基礎となる。しかし，分析は内省によることが多く，網羅的・客観的に分析結果を検証することが困難である。

また，現状では，客観的な分析をするための資料が十分整備されていない。例えば，言語学的な資料としては,『複合動詞資料集』（野村・石井 1987）や『合成語のためのデータベース』（山下 2007）,『複合動詞リスト』（姫野 1999）などが作成されているが，語構成，連接頻度情報といった語自体の情報が主体であり，用例や格要素の情報は収録されていない。自然言語処理の分野でも，形態素解析システムの辞書の中に複合動詞が登録されている。しかし，網羅的に登録されておらず，形態素解析システムでコーパスを解析しても，すぐには複合動詞用の資料として活用できず，解析後に複合動詞を再認定する必要がある。

以上の背景のもと，上記の内省のプロセスを，大量の実例に基づいて行うための複合動詞データベースを構築する。データベースには，複合動詞とその構成動詞を収録する。収録の可否は，用例が一定量収集できるか否かにより決定する。用例は，多様性，および，収集コストを鑑みて，Web から収集する。個々の動詞は，用例，格要素の情報を保持する。また，複合動詞の場合は，語構成の情報をあわせ持つ。

この後の節では，構築する複合動詞データベースの構造の設計について説明した後，Web データから半自動的に複合動詞データベースを構築する方法を示す。さらに，構築された複合動詞データベースの内容を概観する。最後に，複合動詞データベースを使った分析例を示す。

なお，本研究は，国立国語研究所の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」*の一環として行っている。

---

†http://www2.ninjal.ac.jp/masaya
*http://www.ninjal.ac.jp/research/project/c/bunmyaku/

2　データベースの構築

2.1　設計方針

次の三つの設計方針のもと，複合動詞データベースを構築した。

(a) 対象とする複合動詞は，いわゆる「語彙的複合動詞」(影山 1993) とする
(b) 一定数以上の用例を収集可能な複合動詞を収録する
(c) Web データの特性に合わせた，用例・格要素情報の作成を行う

(a) により，対象とする複合動詞を限定する。影山 (1993) では，生成文法の見地から,「動詞（連用形）＋動詞」タイプの複合動詞を，語彙的複合動詞と統語的複合動詞に分類している。このうち，語彙的複合動詞は，語彙部門で派生され，レキシコンに記述されるタイプの複合動詞である。統語的複合動詞は，統語的複合動詞は統語部門で形成される。統語的複合動詞の前項・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）というように補文的な関係を持ち，意味的に透過的な合成が行われる。本研究で統語的複合動詞を扱わないのは，その性質上，前項動詞と複合動詞との格関係が明らかだからである。

(b) は，データベースの利用目的である,「複合動詞と構成動詞との関係を，大量の用例に基づいて分析する」ことを達成するために設けた。このことは，Web の使用実態を反映した複合動詞を収録することにもつながる。

(c) は，通常の書籍にはない，Web 特有の性質に対応するために設けた。例えば，Web データには，同一表現の繰返し（例：掲示板におけるスレッドのタイトル）や，ページ単位での引用など，用例・格要素の頻度を計測する際に「ノイズ」となる要素が存在する。これらの影響を軽減しつつ，用例・格要素を収集する。

2.2　Web コーパスの構築と収録動詞の選定

データベースに収録する複合動詞，構成動詞の選定手順は，次のとおりである。この方法の特徴は，個々の複合動詞，構成動詞ごとに Web コーパスを構築し，収録できる用例の量を確認しつつ，収録動詞を選定していくことである。

(1) 『複合動詞資料集』から，複合動詞の構成要素として多用される動詞上位 10 語を選択し,「種動詞」とする。そして，Baroni・Bernardini (2004) の方法で，それぞれ個別に Web コーパスを作成する。個々の Web コーパスのサイズは，10000 ページ（前項の動詞用に連用形で 5000 ページ，後項の動詞用に終止形で 5000 ページ）である。
(2) 作成した Web コーパス中のデータを形態素解析した後，種動詞＋動詞，動詞＋種動詞のパターンを抽出し，複合動詞候補の頻度表を作る。
(3) 複合動詞候補のうち，一定数以上（今回は頻度 5 以上）出現した複合動詞候補を人手で確認し，個別に Web コーパス（サイズは 2000 ページ）を作成する。さらに，2.3 節の「用例の抽出と格解析」を行ない，最終的にデータベースに収録するかを決定する。
(4) 収録が決定した複合動詞の構成動詞を種動詞として，再帰的に (1)〜(4) を繰り返す。例えば，複合動詞「切り捨てる」の場合，後項動詞の「捨てる」が種動詞となる。

2.3　用例の抽出と格解析

前節までの処理で，動詞ごとの Web コーパスが構築される。これらの Web コーパスを対象に，用例の抽出と格解析を行う。なお，用例の抽出と格解析は，個別の Web コーパスごとに行うものであり，統合した Web コーパスに対して，実施するわけではない。

用例抽出は，収集対象の動詞を含む「文」を単位とする。「文」の区切りは，句読点，空白文字を用いた。収集対象の動詞を含むか否かは，形態素解析した結果に基づいて判断する。用例を 100 例以

上収集できた動詞は収録対象とし，個々の用例に対して構文解析，および，格解析を行う。この結果をもとに，用例に格要素情報（格助詞と格要素のペア）を付与する。なお，形態素解析には JUMAN (ver.6.0)，構文解析・格解析には KNP (ver.3.01) を用いた。

なお，設計方針 (c) に対応するため，収録動詞を決定する際の用例数の計測と，データベースに対する検索処理（3.2 参照）の際に，次の処理を行なっている。

- 用例の出現頻度は，出現ページ数として計測する。ただし，まったく同一の用例は複数のページに出現していたとしても，重複して計測しない。また，データベースへの重複登録も行わない。
- 格要素の名詞の出現頻度も，出現するページ数（以後，出現ページ数）として計測する。例えば，同一の Web ページ内で「畑でトマトを作る」「太郎がトマトを作る」という用例が出現したとしても，ガ格の格要素としての「太郎」は，頻度 1 である。

3 構築されたデータベース

3.1 収録データ

2 節の方法を用いて，複合動詞データベースを構築した。収集した動詞は，複合動詞 3399 語，構成動詞 1075 語である†。各動詞ごとに収集された用例は，複合動詞が平均 1088.4 文（異なりページ数 784.8 ページ），構成動詞が平均 7839.1 文（異なりページ数 2922.8 ページ）となった。複合動詞のうち，用例数が 1000 以上収集できたものは，1839 動詞であった。用例数の分布をヒストグラムにした結果を図 1，2 に示す。

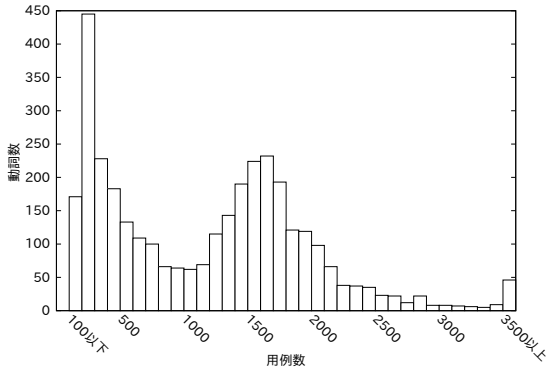

図 1: 収集された用例数の分布（複合動詞）

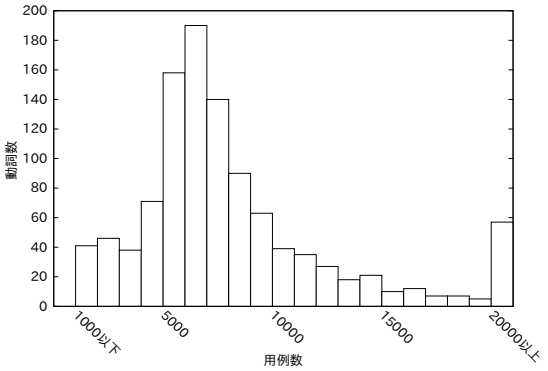

図 2: 収集された用例数の分布（構成動詞）

次に，構成動詞の内訳は，異なりで 999 語となった。前項・後項動詞を個別に見てみると，前項動詞が異なりで 719 語，後項動詞が異なりで 578 語である。使用頻度上位 5 位を，それぞれ表 1，2，3 に示す。

表 1: 構成動詞上位 5 位（前項）

表記	読み	度数
見る	みる	73
引く	ひく	67
取る	とる	60
打つ	うつ	58
突く	つく	48

表 2: 構成動詞上位 5 位（後項）

表記	読み	度数
込む	こむ	228
上げる	あげる	130
出す	だす	122
上がる	あがる	70
取る	とる	65

表 3: 構成動詞上位 5 位（前後）

表記	読み	度数
込む	こむ	231
上げる	あげる	131
出す	だす	126
取る	とる	125
見る	みる	80

†本稿執筆時点では構築途中のため，今後，増減する可能性がある

3.2　データベースの機能

構築したデータベースは，言語研究者を始めとした一般の利用者が手軽に利用できるように，検索用のサイト‡を試験的に公開している。ここでは，その機能の一部を紹介する。

複合動詞検索　読み，もしくは，表記を指定して，複合動詞を検索する。構成動詞を指定した場合は，当該の構成動詞の他，構成動詞を含む複合動詞の一覧が検索される。一覧から詳細に調べたい動詞を選択すると，当該動詞と構成動詞との「重複度」（山口 2012），および，格要素一覧（後述）などが表示される。図 3 は,「言い当てる」を検索した例である。

重複度は，二つの動詞の格要素が重複して使用される度合いを表す。図 3 の例では,「言い当てる」と「言う」の重複度が 46%となっている。これは,「言い当てる」のヲ格の格要素のうち,「言う」のヲ格で使用されている格要素が 46%あることを示している。このように，重複度により，複合動詞と構成動詞との関係を客観的に捉えることができる。

格要素閲覧　検索した動詞の格要素を，格ごとに一覧表にして表示する。図 3（下段の表）が,「探し出す」の格要素一覧である。それぞれの格要素の右には，出現したページ数が付与される。格要素は，ページ数で降順に表示され，格も格要素のページ総数が多い順に左から右に表示される。

複合動詞と構成動詞との関係分析を支援する機能として，構成動詞でも使用される格要素を色分けして表示することができる。また，それぞれの格要素には,（後述の）「用例閲覧」機能用のリンクがあり，容易に用例を閲覧できるようになっている。

用例閲覧　用例を文単位で表示する。各用例には，出典情報として，取得元の Web ページの URL が併記される。

重複度/格パターン

	ヲ	修飾	ガ	デ	ニ	時間	総ページ数
言い当てる	631 p	146 p	36 p	42 p	10 p	11 p	1129 p
言う	46%	90%	100%	21%	70%	64%	3429 p
当てる	44%	85%	75%	57%	70%	64%	2050 p

格要素一覧　（重複：v1　v2　v1,2　off）

ヲ		修飾		ガ		デ		ニ		時間	
こと	59	正確に	34	人	12	見た	17	とき	4	時	4
名前	46	ズバリ	32	者	9	一言	9	こと	4	瞬時	4
未来	27	ずばり	15	それ	6	発	7	前	3	あと	3
本質	23	見事に	13	これ	5	見る	5	夜明け	3		
犯人	20	的確に	7	方	3	聞いた	4				

図 3: データベースの検索結果例

4　活用例：格要素の重複度による複合動詞・構成動詞の分析

4.1　分析のねらい

構築した複合動詞データベースの活用例として，重複度を用いて，複合動詞・構成動詞間の関係を分析してみる。

複合動詞と構成動詞間の関係を記述する場合，従来の分析では，構成動詞の性質がどのように複合動詞に継承されるか，十分に記述されていない。例えば，山本 (1984) では，複合動詞と構成動詞とが格支配構造上の関係を持つか否かによって，複合動詞を 4 種類に分類している。また，由本 (2005) の LCS（Lexical Conceptual Structure）による分析では，構成動詞の LCS が複合動詞の LCS の一

‡http://csd.ninjal.ac.jp/comp/

部に組み込まれる形で記述されている。これらの分析では，構成動詞の性質が複合動詞にそのまま継承されるように見える。

それでは，複合動詞と構成動詞に密接な関係があると考えられる「投げ捨てる」の場合,「投げ捨てる」のヲ格の格要素は,「投げる」のヲ格の格要素としても使えるだろうか？　逆に，前項動詞の意味が希薄な「打ち捨てる」のヲ格の格要素は,「打つ」でもヲ格の格要素とならないだろうか？ この疑問を重複度を用いて検証することが，本節の分析のねらいである。

4.2　分析対象の動詞と重複度

ここでは，問題を簡略化するために，対象とする動詞を，後項動詞に「込む」を持つ複合動詞とし，対象とする格はヲ格とする。また，格解析などの誤りによるノイズを軽減するため，対象とする動詞には，(a) 複合動詞，構成動詞ともに 1000 例以上の用例を持つこと，(b) ヲ格の格要素を 50 例以上持つこと，を条件として加えた。この結果，対象となる複合動詞は，99 語となった。

分析対象の複合動詞の重複度を複合動詞データベースに基づいて計算し，重複度の昇順にプロットした結果を図 4 に示す。

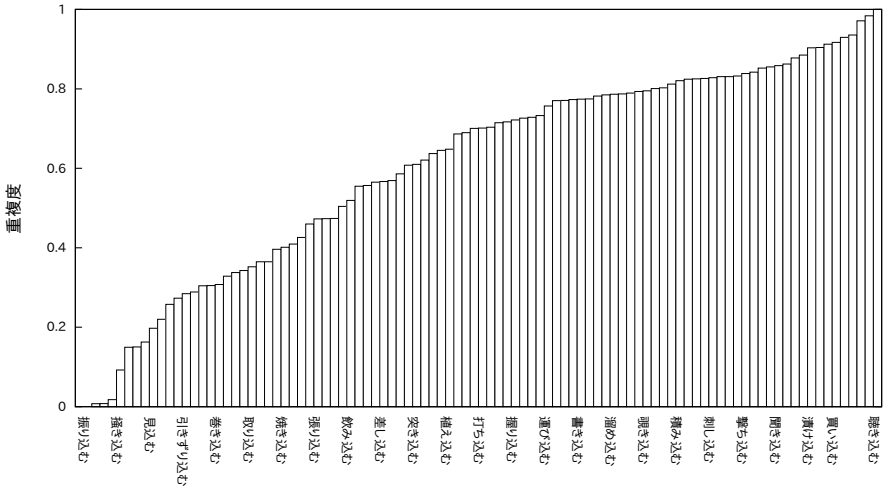

図 4: 重複度の分布

4.3　重複度による複合動詞・構成動詞間の関係の分類

既存の分析を極端に解釈すれば，重複度は，0 と 1 の両端周辺に集中するはずである。しかし，図 4 のとおり，幅広く分布している。この要因を探るために，重複度によって，複合動詞・構成動詞間の関係を次の四つに分類し，考察した。

継承　重複度が 1.0 に近い場合である。この場合，複合動詞の格要素は，構成動詞とほぼ一致する。一致しない格要素については，データ不足が原因と考えられる。今回，重複度 0.9 以上の動詞は 8 例あった。上位の 3 語（括弧内は重複度）は,「聞き込む」(1.0)「着込む」(0.97)「塗り込む」(0.94) である。このうち，一致しなかった格要素を見てみると,「レインウェアを 着込む」「溶剤を 塗り込む」というように，意味的には構成動詞の格要素としても問題ないものだった。このような不一致は，収集する用例数を増やせば，減少するものと考えられる。

別義　継承とは逆に，重複度が 0 に近い場合である。この場合，複合動詞と構成動詞との意味的な関係が少ない，別義と考えられる複合動詞であった。重複度の下位 3 語は,「振り込む」(0.0)「擦り込む」(0.01)「申し込む」(0.01) である。なお，わずかながら一致した格要素は，次のように，限定された格要素だけであった。

- 「傷口にアルカリ性の 墨を 擦り込む」(http://b.z-z.jp/thbbs.cgi/kangofu/331/)
- 「その 旨を フロントまたは管理室に申し込む」(http://www.mm-m.ne.jp/glkumiai/kiyaku_2.html)

派生　継承と別義が混在することにより，重複度が減少する場合である。「織り込む」(0.29) と「追い込む」(0.56) の例を次に示す。いずれの例共に，右側が継承，左側が別義の関係にある。このように，別義に相当する格要素が，重複度の減少の要因となっている。

糸を織り込む ⟺ 糸を織る　　最新の情報を織り込む ⟺ * 情報を織る
魚を追い込む ⟺ 魚を追う　　内閣を総辞職に追い込む ⟺ * 内閣を総辞職に追う

分布変化　格要素の生起確率の分布が，複合動詞と構成動詞で大きく異なる場合である。まず，実際の例として,「流し込む」の格要素を見てみよう。表 4 は，複合動詞となった時に生起確率が上昇した格要素，表 5 は下降した格要素である。前者は上位 10 語のうち，重複度を減少させた 5 語（つまり，頻度 0 だった語）を挙げている。後者は，上位 5 語である。

表 4: 生起率が上昇した格要素

格要素	複合動詞	構成動詞	増減
モルタル	0.014	0.000	0.014
樹脂	0.012	0.000	0.012
ビール	0.012	0.000	0.012
金属	0.011	0.000	0.011
セメント	0.011	0.000	0.011

表 5: 生起率が下降した格要素

格要素	複合動詞	構成動詞	増減
血	0.002	0.068	-0.066
水	0.043	0.110	-0.067
電流	0.009	0.112	-0.103
情報	0.001	0.118	-0.117
涙	0.001	0.219	-0.218

表 4 に挙げた格要素は，今回，構成動詞「流す」のヲ格の格要素としては出現しなかったが，どの格要素も意味的に「流す」の格要素となりうる。これは，重複度の面から見ると,「継承」と同じ状況である。しかし,「継承」の場合と異なるのは，格要素の生起確率に，系統的な分布変化が見受けられる点である。「流し込む」の例で言えば，モルタルなど，何かの材料となるものや，アルコール飲料などの変化が大きかった。これは，複合動詞と構成動詞との関係を記述する上で，重要な情報になる。今後，生起率が下降した格要素（表 5）とあわせて，分析を進める予定である。

5　おわりに

本稿では，Web データに基づいて，日本語複合動詞のデータベースを半自動的に構築する方法を示し，実際に構築した結果を紹介した。さらに，重複度の面から複合動詞と構成動詞の関係分析を行ない，生起確率の分布変化が両者の関係を記述する際に重要になることを示した。現時点で収録されている複合動詞は 3399 語であり，そのうち，用例数が 1000 以上の複合動詞が 1839 語ある。今後は，複合動詞と構成動詞との関係を分析する過程で，データベースの質的な改善を図る予定である。

参考文献

山本清隆 (1984) 複合動詞の格支配，都大論究，Vol.21，pp.32–49

影山太郎 (1993) 文法と語形成，ひつじ書房

由本陽子 (2005)『複合動詞・派生動詞の意味と統語』，ひつじ書房，pp.110–129

野村雅昭，石井正彦 (1987) 複合動詞資料集，科研費特定研究 (1) 言語データの収集と処理の研究

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 研究成果報告書

姫野昌子 (1999)『複合動詞の構造と意味用法』，ひつじ書房，pp.245–260

M. Baroni and S. Bernardini (2004) BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004

山口昌也 (2012) 複合動詞と構成要素動詞の格要素の対応関係分析，言語処理学会第 18 回年次大会予稿集