〈共同研究プロジェクト紹介〉萌芽・発掘型 : 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 複合動詞用例データベースの構築と活用

著者	山口 昌也
雑誌名	国語研プロジェクトレビュー
巻	4
号	1
ページ	61-69
発行年	2013-06
URL	http://doi.org/10.15084/00000732

国語研プロジェクトレビュー
NINJAL Project Review　Vol.4 No.1 pp.61–69（June 2013）

複合動詞用例データベースの構築と活用

Construction of a Database of Japanese Compound Verb Examples and its Applications

山口 昌也（YAMAGUCHI Masaya）

1. はじめに

日本語には，「塗り替える」のように，複数の動詞が結合した複合動詞がある。それでは，複合動詞とその構成動詞の間にはどのような意味的な関係があるのだろうか？　例えば，「塗り替える」と「塗る」の語義の関係を見てみよう。『大辞林』（松村 2006: 1951, 1952）によれば，次のようになっている。

■「塗る」の語義
1. 物の表面に液や塗料，また，ジャム・バターなどをなすりつける。
2. 壁土や漆喰などをなすりつけて，壁や塀などをつくる。
3. （おしろいをつけて）化粧をする。
4. 罪や責任を他人になすりつける。

■「塗り替える」の語義
1. 前に塗ってあったものを改めて，新しく塗り直す。
2. すっかり変える。また，記録などを更新する。

筆者の語感で言えば，「塗る」の語義 1，2 が「塗り替える」の語義 1 に継承されていると感じられる。ただ，ジャムやバターを「塗り替える」ことは，限られた状態でしか考えられない。また，「塗り替える」の語義 2 と「塗る」の語義との間に関係があるとは感じられない。

以上は，筆者の語感でしかなく，筆者の知らないところで，パンのジャムを塗り替える人たちがたくさんいるかもしれない。そこで，より客観的な分析を行うために，複合動詞と構成動詞の用例データベースを作成した。この用例データベースを用いれば，分析対象の動詞の用例や取りうる格要素（例えば，「塗り替える」のヲ格にどのような名詞を取るか）を検索できるだけでなく，複合動詞，構成動詞が共通に取りうる格要素を簡単に調べることができる。

以下，本稿では，複合動詞用例データベースの構築方法，および，その活用例を示す。

2.　複合動詞用例データベースの構築

2.1　概要

複合動詞用例データベースは，複合動詞，および，その構成動詞の用例を収録している。収録対象の複合動詞は，「動詞（連用形）＋動詞」タイプの複合動詞のうち，「語彙的複合動詞」（影山 1993）である。この種の複合動詞は，複合時に意味的な制限が加わったり，まったく別の意味を持ったりするようになる。なお，「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）のような「統語的複合動詞」は，「意味関係は完全に透明かつ合成的」（影山 1993: 78）であるとされているため，基本的に収録対象としない。

収録対象の複合動詞は，Web から収集可能な用例数に基づいて，探索的に決定する。個々の動詞に対して，付与する情報は，次のとおりである。

・複合動詞（表記・読み，語構成情報，用例，格要素情報）
・構成動詞（表記・読み，用例，格要素情報）

用例は，文単位で収録する。格要素情報は，格助詞と格要素のペア，さらにその出現ページ数から構成される。次の例は，複合動詞「聞き出す」の格要素情報である。なお，かっこ内の数字は，出現ページ数を表す。

ヲ格	情報（159）/話（67）/番号（59）/名前（37）/本音（33）/住所（31）/場所（31）/秘密（24）
カラ格	人（15）/本人（11）/相手（9）/者（9）/男（7）/彼女（6）/彼（6）/こちら（4）/口（4）/子供（3）
デ格	電話（7）/中（7）/会（5）
ニ格	人（6）/中（5）/時（4）/前（4）

2.2　構築方法

本構築方法の特徴は，漸進的に用例データベースを構築していくことである。具体的には，巨大な一つのコーパスから用例を採取するのではなく，個々の複合動詞，構成動詞ごとに専用の Web コーパスを構築し，そこから用例を採取する。この方法の利点は，（a）収集できる用例数を調査しつつ，構築を漸進的に進められること，（b）低頻度語であっても，効率的に用例を収集できることである。

複合動詞用例データベースの構築方法を図 1 に示す。構築手順を以下に示す。なお，図中の番号は，手順の番号と対応している。

（1）まず，複合動詞の構成要素になりやすい，「種（たね）」となる構成動詞（以後，「種動詞」）を用意する。今回は，『複合動詞資料集』（野村・石井 1987）から上位 10 語を選択した。
（2）次に，Baroni らの方法（Baroni et al. 2009）を応用して，種動詞に対する Web コーパスを構築する。具体的には，種動詞とランダムな語のペアをキーとして，Web 検索エンジンに与え，得られた URL の Web ページをダウンロードする。ランダムな語をキーに加えているのは，収集する Web ページの偏りを防ぐためである。種動詞は，終止形，連用形の 2 種類用意する。そして，それぞれ 5000 ページずつ収集し，それぞれ独立した Web コーパスとする。終止形で検索するのは，種動詞を後項に持つ複合動詞を発見す

るため，連用形で検索するのは，前項に種動詞を持つ複合動詞を発見するためである。

(3) 構築したWebコーパスを形態素解析[1]したのち，「動詞(連用形)＋動詞」の並びを複合動詞候補として，頻度を計測する。

(4) 得られた複合動詞候補のうち，頻度5以上のものを目視で確認し，複合動詞であれば，複合動詞リストに追加する。

(5) 複合動詞リスト中の複合動詞のWebコーパスを作成する。収集するWebページは，2000ページである。それぞれのWebコーパスを形態素解析し，当該の複合動詞を含む文を抽出する。抽出した文は，格解析[2]，および，同一文削除などのクリーニングをしたのち，用例データベースに追加する。ただし，格要素を一つ以上持つ用例が50例未満の場合は，登録しない。また，登録した場合は，その構成動詞の用例も用例データベースに登録する。

(6) (5) の複合動詞リスト中の複合動詞の構成動詞のうち，種動詞でないほうの構成動詞を種動詞として，(1)～(6) を繰り返す。

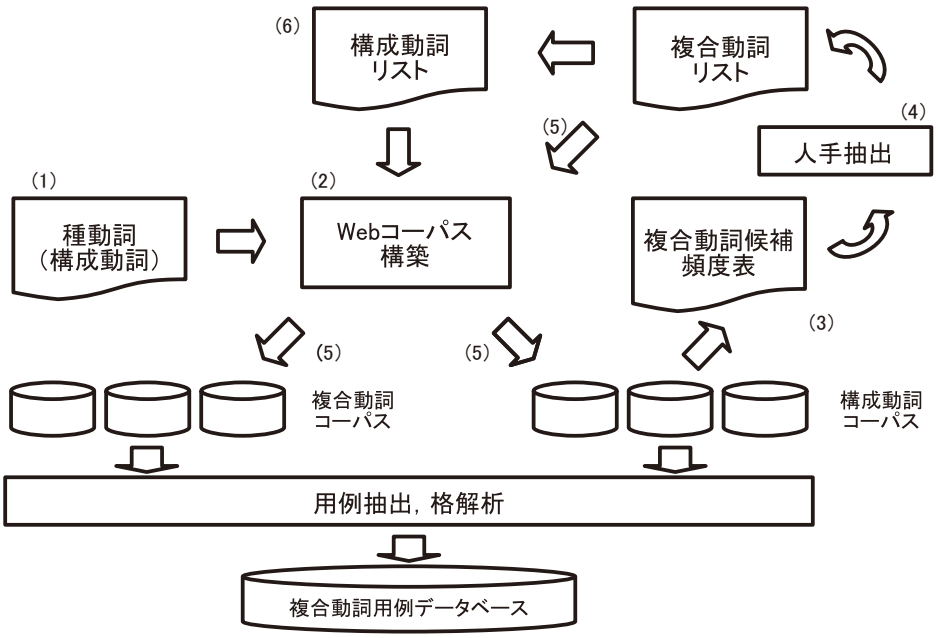

図1　用例データベースの構築方法

[1] 形態素解析システムは，JUMAN（京都大学黒橋・河原研究室，http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN）を利用した。

[2] 格解析には，KNP（京都大学黒橋・河原研究室，http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP）を利用した。

2.3　構築結果

2.2 の構築方法により，複合動詞 3371 語（用例数の中央値 1173 文），構成動詞 963 語（用例数の中央値 5943 文）を収集した。収集した複合動詞の用例数の分布をヒストグラム（図 2）にして示す。横軸は，収集した用例数（用例が出現したページ数の異なり），縦軸は複合動詞の頻度である。

用例数の分布を見ると，高頻度（1250 例周辺）と低頻度（50 例周辺）の二つのピークがある。用例数を考慮すると，前者は定着している複合動詞，後者はまだ定着していないか，特定の領域でしか用いられない複合動詞であると考えられる。

登録した複合動詞の内訳を調べるために，「岩波国語辞典第五版タグ付きコーパス 2004」（岩波書店 2010）の複合動詞と比較してみると，岩波国語辞典の登録語の約 77.2% が収集できていた。図 2 の●は，岩波国語辞典と重複している複合動詞の分布である。この図のとおり，岩波国語辞典の登録語は，高頻度側の分布に類似している。一方，岩波国語辞典のみに登録されていた複合動詞（図 2 の×）は，低頻度語が中心（中央値 74）である。なお，複合動詞用例データベースのみに登録されていた複合動詞は，2114 語である。

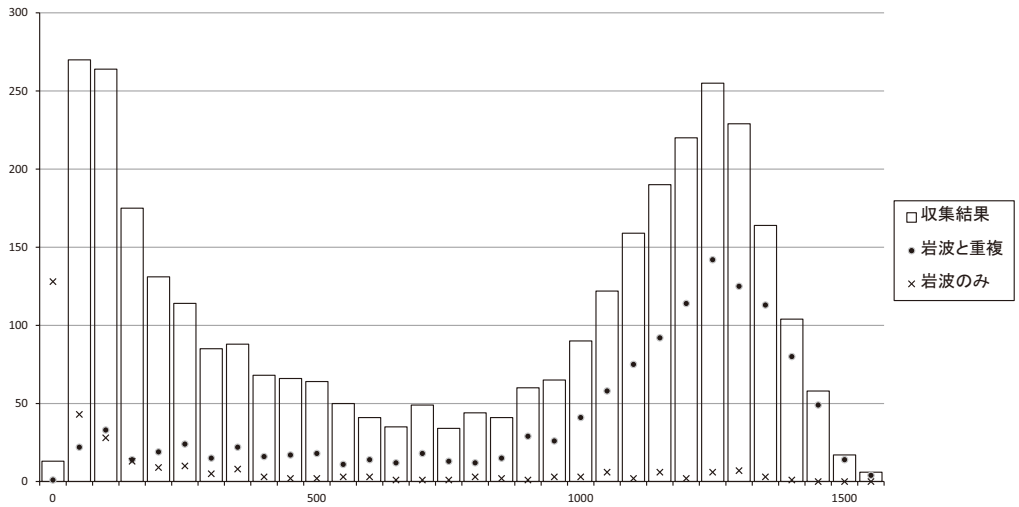

図 2　収集した複合動詞の用例数

3.　データベースの活用

3.1　『Web データに基づく複合動詞用例データベース』の一般公開

構築した用例データベースを一般の利用者も利用できるように，検索システムとともに Web 上に公開した[3]。ここでは，動詞の検索機能と，格要素一覧機能を紹介する。図 3 は，構成動詞「あてる」で検索した例である。左右の表は，それぞれ前項，後項の読みが「あてる」の複合動詞一覧（一部）である。複合動詞（最左列）をクリックすると，後述の格要素一覧が

[3] http://csd.ninjal.ac.jp/comp/

前項一致

		前項		後項	
当てこする	あてこする	当てる	あてる	こする	こする
当て込む	あてこむ	当てる	あてる	込む	こむ
当てつける	あてつける	当てる	あてる	つける	つける
当てはまる	あてはまる	当てる	あてる	はまる	はまる

後項一致

		前項		後項	
射当てる	いあてる	射る	いる	当てる	あてる
言い当てる	いいあてる	言う	いう	当てる	あてる
押し当てる	おしあてる	押す	おす	当てる	あてる
嗅ぎ当てる	かぎあてる	嗅ぐ	かぐ	当てる	あてる

図 3　検索結果の例

表示される。構成動詞の部分のリンクは，当該のキーで再検索するためのものである。

図 4（上）の表は，下側の格要素一覧の概要である。格ごとに格要素の総出現ページ数（格要素一覧の出現ページ数を合計した結果）と，収集した「塗り替える」の用例の総出現ページ数を表示する。また，構成動詞と複合動詞の格関係を分析するために，格ごとの「重複度」を表示する。重複度 OV_i は，複合動詞と構成動詞の格 i が同一の格要素を取る程度を表し，次の式で定義される。なお，E_{ci}，E_{si} はそれぞれ複合動詞，構成動詞の格 i の格要素集合，w_a，w_b は格要素の名詞，n(w) は w の出現頻度を表す。

$$OV_i = \sum_{w_a \in E_{ci} \cap E_{si}} n(w_a) \; / \; \sum_{w_b \in E_{ci}} n(w_b)$$

例えば，図 4（上）の表では，「塗り替える」のヲ格の格要素のうち，出現ページ数で 13% の格要素が「塗る」のヲ格でも用いられることを示している。重複度は，複合動詞の格要素の集合が，構成動詞の格要素の部分集合の場合 1，重複する部分がない場合 0 となる。

「塗り替える（ぬりかえる）」

重複度/格パターン

	ヲ	修飾	ニ	時間	ガ	デ	カラ	総ページ数
塗り替える	993 p	210 p	91 p	90 p	23 p	17 p	6 p	1365 p
塗る	13%	86%	97%	97%	48%	82%	50%	3468 p
替える	8%	56%	8%	90%	30%	0%	0%	2393 p

格要素一覧　（重複：v1　v2　v1,2　off）

ヲ		修飾		ニ		時間		ガ		デ		カラ	
記録	345	大きく	36	色	43	年	53	俺	4	塗装	5	根底	3
歴史	154	大幅に	20	黒	7	毎年	6	人	4	色	5	後	3
地図	75	的に	16	塗装	6	今	5	者	3	そこ	4	これ	3
図	50	また	10	カラー	6	前	4	それ	3	塗料	4		
色	42	一気に	10	ピンク	5	日	4	数	3	一瞬	3		
史	37	ため	9	黄色	5	中	3	韓国	3				
イメージ	23	全て	6	それ	4	今後	3	マーク	3				
外壁	22	どう	6	前	4	後	3	フェイスブック	3				

図 4　重複度と格要素一覧の例

図 4(下）の表は，「塗り替える」の格要素を格ごとに集計し，出現ページ数順で表示した結果である(一部)。それぞれの格要素には，用例閲覧用のリンクがついており，クリックすると当該の格要素を含んだ用例が一覧表示される。また，出現ページ部分が網掛けになっている格要素は，構成動詞(この図の場合は，前項動詞（V1）の「塗る」）でも格要素となりうることを示す。例えば，ヲ格を見ると，「色」「外壁」が網掛けになっており，「塗る」と共通する格要素であることがわかる。これにより，「塗る」の語義 1，2 が「塗り替える」に引き継がれている証拠となる。なお，「塗る」と「塗り替える」の重複度が 13% であることを考えると，「塗り替える」の語義 2 の用例のほうが，語義 1 よりも多いこともわかる。

3.2　重複度による複合動詞と構成動詞間の関係分類

構築した複合動詞用例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類してみる。

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，103 語となった。

分析対象の複合動詞の重複度を複合動詞用例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 5 に示す。

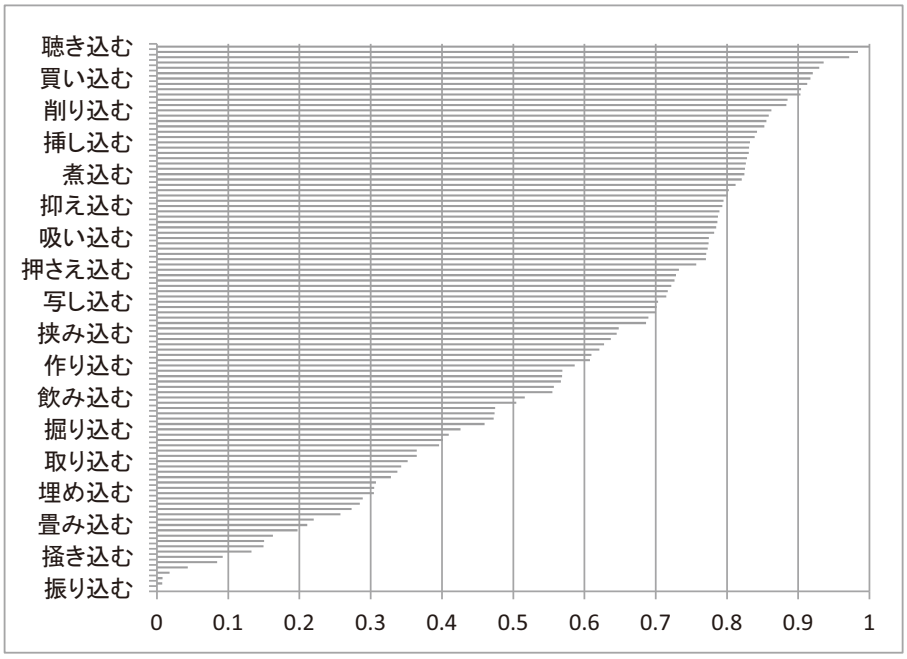

図 5 「〜込む」と前項動詞との重複度

複合動詞と構成動詞間の関係が，関係の有無のみで記述できるとすると，重複度が 0，もしくは，1 付近に集中するはずであるが，この図のとおり，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。今回，重複度 0.9 以上の動詞は 8 例あった。上位の 3 語（かっこ内は重複度）は，「聴き込む」（1.0）「着込む」（0.97）「塗り込む」（0.94）である。

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。

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。どちらも，左側が継承，右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：格要素の生起確率が複合動詞と構成動詞で大きく異なり，複合動詞側では頻出する格要素が，構成動詞側では，ほとんど出現しない場合である。これは，重複率を下げる要因となる。実例として，「流し込む」（0.46）の格要素を見てみよう。生起確率の差で上位 5 位を挙げると，次のようになる（かっこ内の左は複合動詞での頻度，右は構成動詞での頻度）。これらは，別義と異なり，「流す」にとって，不適格な格要素ではないことに注意されたい。

モルタル（19, 0），樹脂（17, 0），ビール（17, 0），金属（16, 0），セメント（15, 0）

以上のように，重複度を用いることにより，複合動詞，構成動詞間の意味的な関係を把握するてがかりになるとともに，分布変化のように，内省では記述するのが難しい現象を明らかにすることができる。

4.　おわりに

本稿では，複合動詞用例データベースの構築方法と構築結果を示した。また，その活用例として，Web で一般公開した検索システムの機能を紹介するとともに，重複度による「～込む」と前項動詞との関係分類を試みた。

●参照文献●

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi and Eros Zanchetta（2009）The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43（3）: 209−226.
岩波書店（2010）岩波国語辞典第五版タグ付きコーパス 2004（CD−ROM）．言語資源協会.
影山太郎（1993）『文法と語形成』東京：ひつじ書房.
松村明（編）（2006）『大辞林』第 3 版．東京：三省堂.
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」研究成果報告書.

《要旨》 本プロジェクトでは，複合的な言語要素の意味記述の一環として，日本語の複合動詞を対象に，複合動詞と構成動詞間の関係記述を行った。本稿では，客観的な関係記述を行うための基礎資料である，複合動詞用例データベースの構築方法を示した。構築の結果は，複合動詞 3371 語（用例数の中央値 1173 文），構成動詞 963 語（用例数の中央値 5943 文）となった。これは岩波国語辞典（第五版）の 77.2% の複合動詞をカバーする。構築したデータベースの活用例として，Web 上に一般公開した検索システムの機能を紹介した。さらに，共通して取りうる格要素の割合に基づき，「〜込む」と前項動詞との関係分類を試み，4 種類の関係を明らかにした。

Abstract: Our project investigated the relationship between Japanese compound verbs and their components as part of an effort to describe compound linguistic elements. In this paper, we present a method of constructing a database of Japanese compound verb examples, which provides basic data for objective description. The constructed database includes 3,371 compound verbs（median number of examples = 1,173）and 963 component verbs（median number of examples = 5,943）. The database covers 77.2% of the compound verbs included in the dictionary *Iwanami Kokugo Jiten*（5th edition）. As an application of the database, we introduce a Web-based search system and show how the relationship between a compound verb ending in *-komu* and its first component verb can be classified into one of four types by using the overlap rate of the case-marked elements that both verbs can take.

山口 昌也（やまぐち・まさや）

国立国語研究所言語資源研究系准教授。博士（工学）（東京農工大学）。東京農工大学助手，国立国語研究所研究員，主任研究員，助教を経て，2010 年 10 月より現職。
主な著書・論文：「相互教授モデルに基づく学習者向け作文支援システムの実現」（共著，『自然言語処理』16（4），2009），「構造化された言語資料に対する全文検索システムの設計と実現」（共著，『自然言語処理』12（4），2005），「前編集結果を利用した前編集自動化規則の獲得」（共著，『情報処理学会論文誌』39（1），1998）.

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
（国立国語研究所　言語資源研究系　准教授）

プロジェクトの概要

文脈情報は，従来から，シソーラスの自動構築，多義語の曖昧性解消など自然言語処理のタスクにおいて利用されてきた。多くの研究では，「類似する文脈に出現する語は意味的にも類似している」という「分布仮説」を前提としており，文脈情報は一種の意味記述として利用されている。本研究プロジェクトでは，単語周辺の文脈情報から，複合的な言語要素（例：複合動詞）の意味記述（文脈情報）を合成的に導出する理論の確立を目指し，（1）（個々の）単語周辺の文脈情報と，複合的に用いられたときの文脈情報との関係の解明，（2）文脈情報の表現方法などを含めた分布仮説の検証，（3）自然言語処理結果の言語学的観点からの検証，を行う。