用例に基づく複合動詞の構造分析と教育への応用

山口昌也　(国立国語研究所言語資源研究系) †
井上 優　(麗澤大学外国語学部)
柏野和佳子　(国立国語研究所言語資源研究系)
北村雅則　(名古屋学院大学商学部)
白井清昭　(北陸先端科学技術大学院大学情報科学研究科)
千葉庄寿　(麗澤大学外国語学部)

Analysis of Japanese Compound Verb Based on Examples and its Application to Education

Masaya YAMAGUCHI (Dept. Corpus Studies, NINJAL)
Masaru INOUE (Faculty of Foreign Studies, Reitaku University)
Wakako KASHINO (Dept. Corpus Studies, NINJAL)
Masanori KITAMURA (Faculty of Commerce, Nagoya Gakuin University)
Kiyoaki SHIRAI (School of Information Science, JAIST)
Shouju CHIBA (Faculty of Foreign Studies, Reitaku University)

1　はじめに

本稿では，大量の用例に基づいて，複合動詞とその構成動詞について，格要素間の関係を分析する。また，分析用のデータ，および，分析結果を日本語教育へ応用する計画についても述べる。

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。本研究は複合的言語要素として，複合動詞を扱ったものである。本稿では，日本語教育への応用について触れるが，プロジェクトとしては，国語辞典編集，語義の自動分類，テンス・アスペクト研究など，国語学，言語学，自然言語処理などと連携させつつ，研究を進めている。

2　複合動詞の構造分析

2.1　概要

本稿では，日本語の複合動詞の格要素に関して，構成する動詞の格要素との関係を分析する。具体的には，(1) 複合動詞と構成動詞の格要素集合が重複する度合いを実際の用例に基づいて計算し，(2) 格要素集合の重複度と複合動詞・構成動詞の格支配構造の関係について考察する。なお，本稿で扱う複合動詞は,「語彙的複合動詞」（影山 1993）である。

日本語の複合動詞と構成動詞との関係分析として，山本 (1984) の格支配構造による分析がある。この研究によれば，複合動詞 (Vc) は，前項動詞 (V1)，後項動詞 (V2) との格支配構造の関係に基づいて，次の四つに分類できるとしている。

- I 類：V1，V2 どちらも Vc の格要素と格支配関係を有するもの (例：「投げ捨てる」「叩き切る」)
- II 類：V1 だけが Vc の格要素と格支配関係を有するもの (例：「見上げる」「書き込む」)
- III 類：V2 だけ Vc の格要素と格支配関係を有するもの (例：「打ち重なる」「引き起こす」)
- IV 類：V1,V2 どちらも Vc の格要素と格支配関係を有しないもの (例：「繰り返す」「取り組む」)

このうち，格支配関係を有する場合，格要素の名詞は複合動詞，構成動詞のいずれの文中でも適格である。例えば，I 類の用例 E1 は，E1a，E1b のように，V1, V2 の用例を作ることができる。

†masaya@ninjal.ac.jp

- (E1) 太郎が煙草を投げ捨てる
- (E1a) 太郎が煙草を投げる
- (E1b) 太郎が煙草を捨てる

このように複合動詞と構成動詞が格支配構造上の関係を有する場合，格要素の名詞も対応関係を有する。したがって，理論上は，複合動詞の格要素の集合は，構成動詞の格要素の集合の部分集合となるはずである。

そこで，本稿では，複合動詞と構成動詞の格要素集合の重複度を調査し，複合動詞・構成動詞の格支配構造における関係の有無が，格要素集合の重複度とどのように関係しているのかを分析する。

2.2 格要素の重複度

本稿では，文における重要性や，格ごとの出現頻度を考慮し，他動詞の場合はヲ格，自動詞の場合はガ格の格要素の重複度を計算する。

格要素の重複度 OV_i は，複合動詞の格 i が取り得る格要素集合 E_{ci} を基準とし，それらが構成動詞の格 i の格要素集合 E_{si} と重複する割合を表す。定義は，次のとおりである。なお，w_a，w_b は格要素の名詞，$n(w)$ は w の出現ページ数を表す。

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

2.3 分析データの構築

2.2 節の分析を行うには，特定の分野に偏らない，大量の用例とそれを格解析した結果が必要となる。そこで，本研究では Web から用例を収集することにした。収集手順は，次のとおりである。

(1) 『複合動詞資料集』(野村・石井 1987) 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」とする。そして，それぞれ 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) を Baroni(2006) の方法で収集する。
(2) 収集した Web ページを形態素解析した後,「種」動詞を含む動詞の連続を抽出し，複合動詞候補とする。そのうち，50 ページ以上に出現した候補を目視確認し，分析対象の複合動詞とする。
(3) 分析対象の複合動詞に対して，Baroni(2006) の方法で 2000 ページの Web ページを収集する。
(4) 収集した Web ページを形態素解析した後，収集対象の複合動詞を含む文を抽出し，構文解析，および，格解析を行う。なお，格解析には，KNP (ver.3.01, http://nlp.ist.i.kyoto-u.ac.jp/) を利用した。
(5) 収集した複合動詞の構成動詞を種として，再帰的に 1〜4 を繰り返す。

以上の手順で複合動詞 783 語，構成動詞 194 語の用例を収集した（原稿執筆時)。平均用例数は，それぞれ 1312.5，14078.4 例である。

2.4 実験

構築した複合動詞のうち，次の条件を満たす複合動詞をランダムに 100 個抽出し，分析対象とした。

- 複合動詞の用例が 1000 個以上収集されていること。また，構成動詞の用例が前項・後項双方とも 2000 個以上収集されていること
- $\sum_{w \in E_{ci}} n(w) \geq 50$，$\sum_{w \in E_{si}} n(w) \geq 50$ であること。ただし，収集した Web ページに出現する割合が，複合動詞の場合，0.25% 未満，構成動詞の場合，0.05% 未満の名詞は除外する。

表 1: 複合動詞の内訳

分類	複合動詞数	構成動詞数	一致率 (%)
I 類	39	48	81.0
II 類	24	30	80.0
III 類	21	26	70.0
IV 類	16	21	64.0
全体	100	80	76.5

以上の条件を満たす複合動詞 100 個に対して，複合動詞と構成動詞との格支配構造上の関係の有無を人手で与えた。山本 (1984) の 4 分類で集計した結果を表 1 に示す。なお，複合動詞が多義の場合，いずれかの語義で格支配の関係が認められれば，関係ありとしている。

表 1 の複合動詞を対象に，人手による格支配関係の判別結果と重複率との関係を見てみる。図 1 に関係のある場合の重複度 ($\mu = 61.1, \sigma = 23.7$)，図 2 に関係のない場合の重複率 ($\mu = 31.8, \sigma = 23.8$) をヒストグラムとして示す。横軸は重複度，縦軸は頻度である。

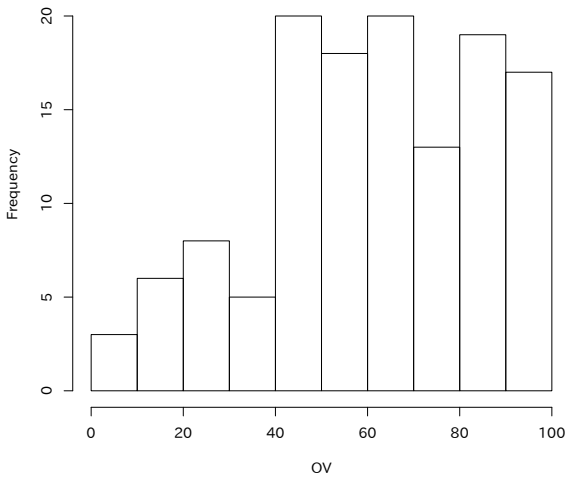

図 1: 重複度 (格支配関係あり)

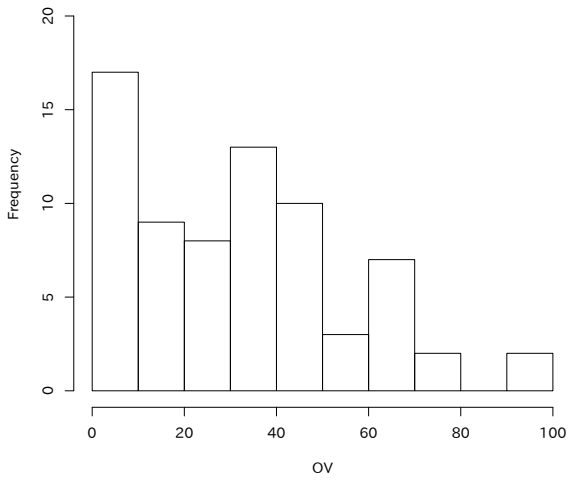

図 2: 重複度 (格支配関係なし)

2.5 考察

まず，人手で付与した格支配関係の有無と重複度による判別結果との一致率を求める。ここでは，閾値 t を設け，$OV_i \geq t$ のとき，格支配関係があると判定するものとする。閾値 t は，0.5 刻みで変化させ，人手で付与した格支配関係の有無との一致率が最大になる値を求めたところ，$t = 38.5(\%)$ となった。表 1 の「一致率」欄に結果を示す。

以上のことより，一致率の面からは，重複度 38.5%に格支配関係の有無の境界がある。この結果や図 1 では，理論とは異なり，格支配関係ありの場合も，重複度の低い場合が存在する。そこで，図 1 の閾値以下の複合動詞を見てみると，最も大きな原因は複合動詞の多義性にあった。例えば,「過ぎ去る」と「去る」のガ格の格要素集合の重複を求めると,「嵐」「台風」「ブーム」などは「去る」の用例と重複するに対して,「時間」「 1 年」などは重複しない。この二つの名詞のグループは，大辞林 (松村 2006) では，二つの語義に対応している。

また，逆に格支配関係なしの場合も，重複度が高い複合動詞が存在する。図 2 の閾値以上の複合動詞を調べてみると，構成動詞の格要素が過剰に重複することが主な原因だった。例えば,「取り扱う」の前項動詞「取る」は，接頭辞的に用いられているため，格支配関係はない。しかし,「取り扱う」のヲ格の格要素集合に「商品」「製品」など,「取る」と共通する名詞が多数含まれ，重複度が高くなる。

3　日本語教育への応用

日本語学習者にとって複合動詞の習得が困難なことは，従来より指摘 (松田 2002 など) されている。そこで，上記の研究成果を日本語教育に応用することを計画している。

その試みの一つとして，収集した用例を検索するシステムを，試験的に全文検索システム『ひまわり』(http://www2.ninjal.ac.jp/lrc) で実現した。実行例を図 3 に示す。この図のとおり，用例を検索し，格要素の一覧などを閲覧することができる。今後，格要素の重複度を応用して，複合動詞と構成動詞との関連を示す仕組みを導入する予定である。

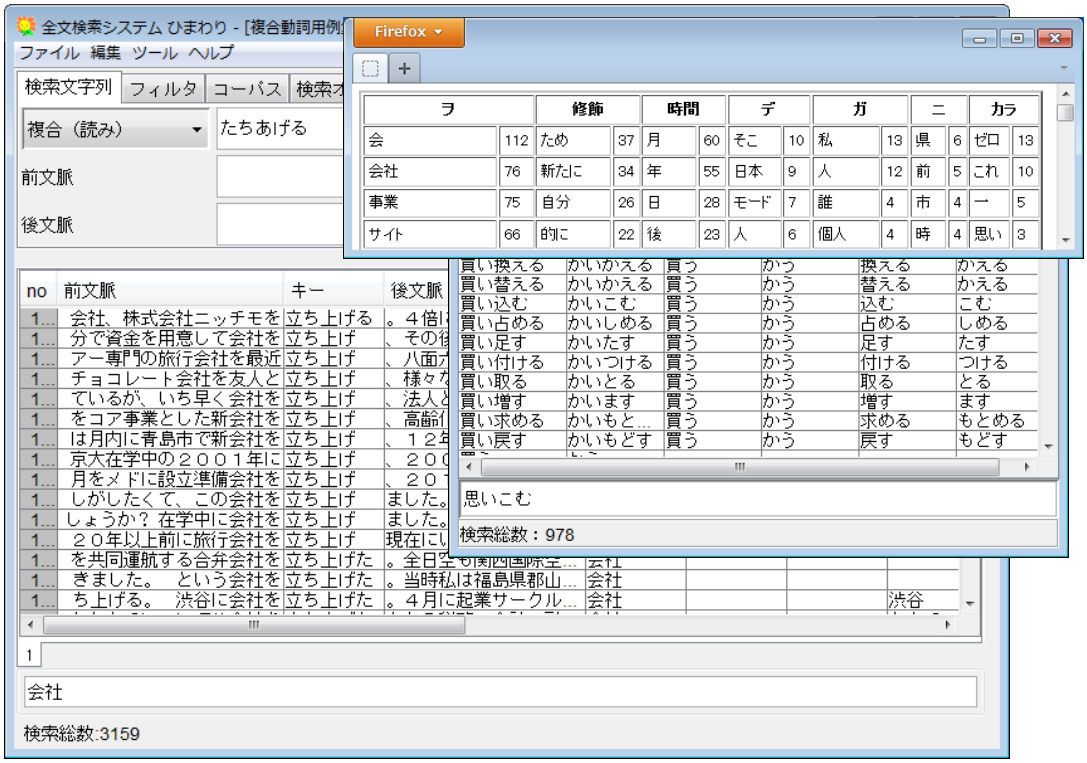

図 3: 全文検索システム『ひまわり』での実現例

4　おわりに

本稿では，日本語の複合動詞の格要素に関して，構成する動詞の格要素との関係を分析するとともに，その応用例として，日本語教育向けの複合動詞用例検索システムを示した。

参考文献

影山太郎 (1993) 文法と語形成，ひつじ書房

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

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

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

松田文子 (2002) 複合動詞研究の概観とその展望 —日本語教育の視点からの考察—，言語文化と日本語教育 増刊特集号，pp.170–184

松村 明 (編) (2006) 大辞林第３版，三省堂