複合動詞と構成要素動詞の格要素の対応関係分析

山口 昌也
国立国語研究所
masaya@ninjal.ac.jp

1. はじめに

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。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) 格要素集合の重複率と複合動詞・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,「生み出す」のような「動詞（連用形）＋動詞」タイプの複合動詞である。

日本語の複合動詞と構成動詞との関係分析として，山本 [1] 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。

- (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

一方，III 類の用例 E2 であれば，V1 の用例 E2a は不適格であり，V2 の用例 E2b は適格である。

- (E2) 災難が打ち重なる
- (E2a) *災難が打つ
- (E2b) 災難が重なる

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。

本稿では，大量の用例を用いて，複合動詞と構成動詞の格要素集合間に上記のような関係が成り立つか実験する。この際，複合動詞，構成動詞の多義性，誤用，解析誤りなどの要因で理論どおりの結果が得られないことが予想される。そこで，複合動詞・構成動詞の格支配構造における関係の有無が，格要素集合の重複率とどのように関係しているのか，実験結果に基づいて考察する。

2. 格要素の重複率

2.1 分析対象の複合動詞の種類

格要素の重複率について述べる前に，本稿で対象とする複合動詞の種類を限定しておく。

日本語の複合動詞については，言語学，日本語学を中心として，多くの研究がなされており，前述の「動詞（連用形）＋動詞」タイプの複合動詞にさらなる分析が加えられている（[2, 3] など）。本研究では，影山 [2] が提案している「語彙的複合動詞」「統語的複合動詞」のうち，語彙的複合動詞を分析対象とする。

[2] の分類は，生成文法中のどの部門で記述されるか基づいている。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。語彙的複合動詞の例として,「使い回す」「出し抜く」「飛び散る」「踏ん張る」を挙げる。これらは，前項・後項動詞の合成時に，意味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」）。

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，意味的に透過的な合成が行われる。

2.2 格要素の重複率

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。

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

なお，w_a，w_b は格要素の名詞，$n(w)$ は w の出現頻度を表す。

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，重複率を計算する格を次のように定める。

- 他動詞の場合： ヲ格
- 自動詞の場合： ガ格

3. 分析データの構築

3.1 構築方法の概要

本節では，分析データの構築方法について述べる。

2.2 節の分析を行うには，特定の分野に偏らない，大量の用例を用意するとともに，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。目標収集量は，複合動詞が 1000，構成動詞が 5000 である。構成動詞の用例数を多くしたのは，構成動詞は多数の格支配構造を持つことが多く，構成動詞の用例を収集しても，複合動詞と同じ格支配構造で用いられる用例は，収集した用例の一部にしかならないからである。

分析データの例として，複合動詞「聞き出す」の格解析結果を次に示す（括弧内の数字は出現頻度）。

ヲ　情報 (166)/話 (68)/番号 (60)/名前 (39)/
　　本音 (33)/住所 (31)/場所 (31)/秘密 (24)
カラ 人 (15)/本人 (11)/相手 (9)/者 (9)/男 (8)/
　　彼女 (6)/彼 (6)/こちら (5)/口 (5)/子供 (4)
デ　電話 (7)/中 (7)/会 (5)
ニ　人 (6)/中 (5)/時 (4)/前 (4)

収集方法の概要は，次のとおりである。この後の節で詳細を説明する。

(1) 分析対象の複合動詞の決定
(2) Web 検索エンジンによる用例の収集
(3) 用例の抽出と格解析

3.2 分析対象の複合動詞の決定

複合動詞に関する資料として，国語辞典や『複合動詞資料集』[5] がある。ただし，前者は網羅的に収録しているわけではないこと，後者は調査年代が古いことから，用例と同様に Web データで漸進的に調査・収集することにした。

具体的な手順は，次のとおりである。

(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」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 3.3 節の方法で収集する。
(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。
(3) 複合動詞候補のうち，一定量以上の Web ページページ [1]で出現する複合動詞候補を目視で確認し，分析対象の複合動詞とする。
(4) 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。

3.3 Web 検索エンジンによる用例の収集

分析対象の複合動詞とその構成動詞の用例を収集する方法として，Baroni らの手法 [4] を応用した。[4] は，ランダムなキーワードを Web 検索エンジンに与え，多様な Web ページからなる Web コーパスを構築する手法である。

本研究では，個別の動詞ごとに一定量以上の用例を収集する必要がある。そこで，巨大な Web コーパスを構築するのではなく，分析対象の複合動詞，構成動詞ごとに Web ページを収集した。量は前述の収集目標を考慮して，複合動詞が 2000 ページ，構成動詞が 5000 ページである。[4] を適用する際は，ランダムなキーワードに加えて，収集対象の動詞をキーワードに付与することにより，収集される Web ページに確実に収集対象の動詞が含まれるようにした。

3.4 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，統合した一つの大きな「Web コーパス」としては実施しない。

用例は分析対象の動詞を含む「文」とする。「文」の区切りは，句読点，空白文字を用いた。用例の抽出は，すべての文の形態素解析を行い，対象の動詞が含まれる文を抽出することにより行う。

次に，抽出された用例に対して構文解析，および，格解析を行う。これにより，用例ごとに格解析結果が付与された状態になる。形態素解析は JUMAN (ver.6.0)，構文解析・格解析は KNP (ver.3.01) を利用した [2]。

なお，上記の処理の過程では，分析に対するノイズ（例えば，Web データに頻出する，ページの全文引用

[1] 今回は，50 ページ以上とした。
[2] http://nlp.ist.i.kyoto-u.ac.jp/

など）に対処するために，次の対策を行っている。

- 格要素の名詞の出現頻度は，出現するページ数として計測する。つまり，同一 Web ページに「太郎がカレーを作る」「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 となる。
- 抽出された用例のうち，重複する用例は削除する。これにより，Web ページ A と B に存在する同一用例の片方は削除する。

3.5 構築結果

前節の方法で，複合動詞，および，構成動詞の用例を収集した。収集結果を表 1 に示す。ただし，前述のとおり，用例の収集は，漸進的に行っているため，本稿執筆時点でデータを用いた[3)]。

表 1: 用例の収集結果

	動詞数異なり	平均用例数
複合動詞	783	1312.5
構成動詞	194	14078.4

4. 実験

4.1 実験条件

3 節で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。

- 用例が 1000 個以上収集されていること
- 認識対象の格の格要素頻度の総数が 50 以上であること
- 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

また，誤解析・誤用などのノイズを減らすために，格要素の名詞の出現頻度が低いものは，分析データから削除した。具体的には，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% 以上，構成動詞の場合，0.05% 以上の名詞だけを用いた。

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。[1] の 4 分類で集計した結果を表 2 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。

4.2 重複率の分布

表 2 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率 ($\mu = 61.1, \sigma = 23.7$)，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。横軸は重複率，縦軸は頻度である。

[3)] 『複合動詞資料集用例集』[5] によれば，国語事典などの辞書に収録されている複合動詞の異なりは，2761 であるとされている。

表 2: 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

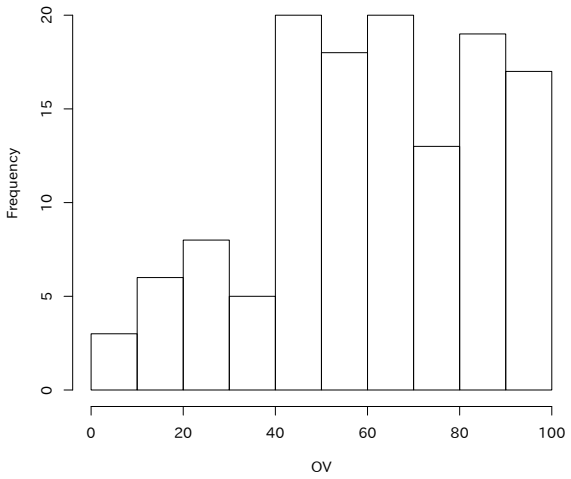

図 1: 重複度 (格支配関係あり)

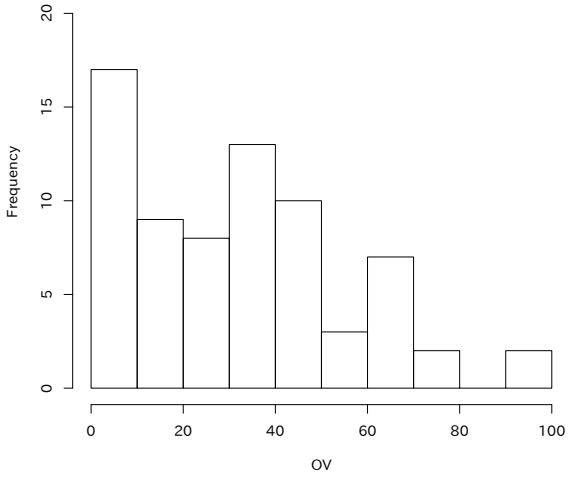

図 2: 重複度 (格支配関係なし)

4.3 一致率

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 2 の「一致率」欄に結果を示す。全体の一致率は，76.5%であった。なお，重複率の計算は前項，後項動詞，別々に行い，一致率も独立して算出している。

5. 考察

本節では，人手による格支配関係の判別結果と重複度による判別結果に差分があるところを中心に，その差異が発生した原因を考察する。

5.1 誤って格支配関係なしと判定される場合

まず，重複率により格支配関係を判別したときに，誤って格支配関係なしと判定される場合について見てみる。この場合は，重複率が低く評価されるために，人手での判断との不一致が生じる。図 1 では，OV_i が閾値の 38.5 より小さい部分に相当する。

実例を分析したところ，(a) 複合動詞の多義性，(b) 構成動詞の用例の少なさ，(c) 誤解析が主な原因であった。ここでは，(a) について詳しく論じる。

(a) の具体例として,「切り開く」の前項動詞「切る」との関係の例を挙げる。人手による判定では,「袋を切り開く」「袋を切る」と言えることから，関係ありとして評価した。しかし，収集した「切り開く」のヲ格の格要素は，次のように，大部分が「新しい道をつける」を意味するものであった。この場合,「切る」に対応する語義はなく，重複率は，2.2%となった。

未来 (206)/道 (204)/時代 (108)/人生 (91)/運命 (59)/世界 (29)/展望 (27) 境地 (25)/性 (20)/将来 (20)/明日 (18)

これらの格要素は,「袋を切り開く」の格支配構造とは異なるので，関係がないと判別されたことは正しい。これらを正しくグループ化することができれば，複数の格フレームを認識する際の手がかりなる可能性がある。

同様の現象は，人手での判断との不一致に影響しないところでもいくつか確認できた。例えば,「過ぎ去る」と「去る」のガ格で重複する格要素の名詞は，嵐 (197)/台風 (48)/ブーム (11) であった。一方，重複せずに,「過ぎ去る」側だけに出現したのが，時間 (75)/日 (24)/時 (23)/年 (23)/日々(17) などであった。この二つの格要素のグループは，国語辞典の大辞林 [6] において,「過ぎ去る」の二つの語義に対応する。

5.2 誤って格支配関係ありと判定される場合

この場合は，前節とは逆に，重複率が高く評価されるために，人手での判断との不一致が生じる。図 2 では，OV_i が閾値の 38.5 以上となる部分に相当する。

この現象が発生する主な原因は，複合動詞の格要素集合が構成動詞の格要素集合と過剰に重複するためであった。例えば,「取り扱う」と前項動詞「取る」の例を見てみる。次に示すのが,「取り扱う」のヲ格の格要素である。なお，* は,「取る」のヲ格の格要素と重複していることを示す。

*情報 (67)/*商品 (54)/*物 (35)/問題 (30)/*製品 (23)/*品 (22)/全般 (15)/*食品 (13)/事務 (12)/旅行 (12)/用品 (11)

この場合,「取り扱う」中の「取る」は接頭辞的に用いられており，格支配の関係はないと判断した。しかし，ヲ格単独で比較すると，上記のように重複する名詞が出てくる。類似の現象は,「思い出す」「感じ取る」「売り上げる」(すべて後項動詞との重複) などでも見受けられた。この現象は,（前節の）重複率が低く評価される場合と異なり，格支配関係の有無の判別にとって誤りとなる。したがって，今後，複数の格での重複度を計算するなどの方法を検討したい。

6. 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した。その結果，(1) 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は 76.5%であること，(2) 人手の判断結果との差異を分析し，複合動詞の格フレーム類別への応用の可能性，および，構成動詞の格要素との過剰な重複の問題を示した。

参考文献

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

[2] 影山太郎：文法と語形成，ひつじ書房 (1993)

[3] 姫野昌子：『複合動詞の構造と意味用法』，ひつじ書房 (1999)

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

[5] 野村雅昭，石井正彦：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 (1987)

[6] 大辞林，三省堂第 3 版 (2006)